cancers

Opinion

Using Electronic Medical Records to Identify Potentially Eligible Study Subjects for Lung Cancer Screening with Biomarkers

Lamorna Brown *, Utkarsh Agrawal and Frank Sullivan

School of Medicine, University of St Andrews, St Andrews KY16 9AJ, UK; ua1@st-andrews.ac.uk (U.A.); fms20@st-andrews.ac.uk (F.S.)
* Correspondence: lb300@st-andrews.ac.uk; Tel.: +44-7824793243

Simple Summary: Recent cancer screening trials have found that using low-dose computed tomography (LDCT), compared to chest radiography, resulted in a significant reduction in lung cancer mortality. To effectively carry out this intervention, individuals at a high risk of developing lung cancer are targeted. However, accurately identifying and retaining these groups can be challenging. As electronic medical records (EMRs) contain important demographic and clinical information, they could be used to accurately identify subjects for screening. To determine whether EMRs can be used for this purpose, this paper examines the evidence around the use of EMRs in screening trials and the information contained in them that could be used to aid researchers in identifying eligible subjects.

Abstract: Lung cancer screening trials using low-dose computed tomography (LDCT) show reduced late-stage diagnosis and mortality rates. These trials have identified high-risk groups that would benefit from screening. However, these sub-populations can be difficult to access and retain in trials. Implementation of national screening programmes further suggests that there is poor uptake in eligible populations. A new approach to participant selection may be more effective. Electronic medical records (EMRs) are a viable alternative to population-based or health registries, as they contain detailed clinical and demographic information. Trials have identified that e-screening using EMRs has improved trial retention and eligible subject identification. As such, this paper argues for greater use of EMRs in trial recruitment and screening programmes. Moreover, this opinion paper explores the current issues in and approaches to lung cancer screening, whether records can be used to identify eligible subjects for screening and the challenges that researchers face when using EMR data.

Keywords: cancer; screening; smoking; electronic records

Citation: Brown, L.; Agrawal, U.; Sullivan, F. Using Electronic Medical Records to Identify Potentially Eligible Study Subjects for Lung Cancer Screening with Biomarkers. *Cancers* **2021**, *13*, 5449. https://doi.org/10.3390/cancers13215449

Academic Editors: Nandini Dey and Pradip De

Received: 6 September 2021
Accepted: 19 October 2021
Published: 29 October 2021

1. Introduction

Lung cancer remains one of the most aggressive and frequently diagnosed cancers in the UK [1]. Mortality rates for the disease remain high, at 21% for both males and females, making it the most common cause of cancer-related death [2]. As late-stage lung cancer (i.e., stage III/IV) is less susceptible to curative medical interventions, such as surgical resection, there is a low survival rate for individuals diagnosed at these stages (2–3%) [2]. The majority of lung cancer cases are diagnosed with late-stage cancer, leading to overall low survival rates at 1 (40%) and 5 years (16%) post-diagnosis [1,3,4].

To reduce late-stage diagnosis, lung cancer screening using low-dose computed tomography (LDCT) has been recommended [5]. Screening trials using LDCT, compared to usual care (i.e., chest X-rays), have provided evidence of a significant mortality benefit.

Trials such as the NLST, NELSON and UK Lung Cancer Screening Trial found those undergoing LDCT scans had a reduced probability of dying from lung cancer [6–8]. 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 [9].

While these trials support the use of LDCT in screening programmes to identify lung cancer, there are practical barriers that can reduce participant engagement, limiting the effectiveness of interventions [10]. These practical barriers include difficulties accessing target groups and identifying patients that fit screening inclusion criteria [10–12]. However, electronic medical records (EMRs) contain important clinical and demographic information that can reduce and resolve these issues [13].

This paper covers the current issues in and approaches to lung cancer screening and appraises the methods used and evidence for the effectiveness and appropriateness of using electronic medical records as a way of identifying those at high risk of developing cancer.

Defining high-risk groups for lung cancer screening is an ongoing challenge. Age, occupation, family history, some respiratory conditions (particularly emphysema) and environmental factors such as air pollution and radon exposure are important risk factors for lung cancer [14,15]. The strongest determinant of lung cancer, however, is smoking, with over 70% of cases in the UK linked to smoking [16,17]. As a result, smoking status has been used to identify eligible participants for lung cancer trials. In this article, we consider an important characteristic of high-risk groups to be whether they are current smokers, and thus papers which report on the recording of smoking in EMRs in order to identify eligible subjects are included in this article. Other health, sociodemographic and environmental risk factors for lung cancer that appear in EMRs are also examined.

2. Issues and Approaches to Current Lung Cancer Screening Programmes

Lung cancer screening programmes use a targeted approach, whereby those most at risk, and thus most likely to benefit from screening, are eligible for inclusion. Trials such as the NELSON and NLST use patient self-declared age and the number of pack years as bases for inclusion using a questionnaire [6,18]. Trials utilising risk models to identify high-risk groups have provided further risk factors to consider for screening criteria, such as family history of lung cancer and respiratory diseases [19]. The use of these models for participant selection has led to lower numbers of individuals eligible for selection but enabled greater prevention of lung cancer death in trials [15,20].

Despite progress in the identification of high-risk individuals, low participation and retention rates can hinder the effectiveness of interventions. Table 1 presents the approach response rates, methods of recruitment and percentage of respondents randomised for some of the major European lung cancer screening trials. Previous lung cancer trials have had approach response rates (i.e., the proportion of individuals who responded when approached) between 23 and 52% [21]. To improve these rates, the barriers and issues around lung cancer screening implementation must be explored.

Table 1. The recruitment strategies, numbers of subjects approached, numbers of respondents and the percentage of respondents randomised in major European lung cancer screening trials.

Lung Cancer Screening Trial	Recruitment Period	Number of Subjects Approached	Number of Subjects That Responded	Approach Response Rate	Number of Eligible Subjects That Consented	% of Respondents Randomised	Method of Recruitment
NELSON [6]	2003–2006	606,409	150,920	24.9%	15,822	10.5%	Direct mail
ITALUNG [22]	2004–2006	71,232	17,055	23.9%	3206	18.8%	Direct mail

LUSI [23]	2007–2011	292,440	95,797	32.8%	4052	4.2%	Direct mail and mass media
NLST [18]	2002–2004	n/a	53,454	n/a	52,486	n/a	Direct mail, mass media and outreach
UKLS [8]	2011–2014	247,354	98,746	39.9%	4061	4.1%	Direct mail
LSUT [24]	2015–2017	2012	1058	52.6%	770	72.8%	Direct mail
LHC Manchester [25]	2016–2018	16,402	2827	17.2%	1384	49.0%	Searched GP records to send direct mail invitations
LHC Liverpool [26]	2016–2018	11,526	4566	39.6%	1318	28.9%	Searched GP records to send direct mail invitations
ECLS [9]	2013–2016	77,077	18,657	24.2%	12,209	65.4%	Searched GP records to send direct mail invitations, mass media and outreach

There are both participant- and provider-related barriers to lung cancer screening engagement. The UK Lung Cancer Screening Pilot Trial identified participant demographic factors associated with a reduced likelihood of participation. It was found that those who were female, older, current smokers and from a lower socioeconomic group were less likely to participate [27]. Further, there are both emotional and practical barriers to participation. Practical barriers such as a participant's state of health and emotional barriers such as fear of screening and information avoidance are cited as reasons for non-participation by eligible subjects [27–29]. The stigma associated with lung cancer may also act as a barrier for both participants and providers [30,31]. Patients with lung cancer report feeling more stigmatised by themselves and others compared to individuals with cancers such as breast, cervical and skin cancer, as there is a perception that they have brought the illness upon themselves by smoking [32]. This can delay individuals seeking help and receiving timely investigation and treatment, which can have a detrimental effect on patient outcomes [33]. Stigma is also associated with reduced levels of screening uptake [34].

The significant barriers that providers face relate to identifying and recruiting eligible subjects. Previous lung cancer screening trials identified subjects through population-based registries [21]. Information that could aid in the identification of high-risk groups may not be present in these registries. Additionally, the information that is present may not be accurate and, as a result, researchers risk contacting individuals who do not meet trial eligibility criteria. Trials that use electronic medical records (EMRs) for identifying subjects have shown that both identification and uptake can match those of trials that have utilised population registries. The LHC Liverpool study utilised EMRs to search for eligible subjects before contacting them; this targeted approach resulted in the trial obtaining one of the highest approach response proportions out of recent lung cancer screening trials (40%) [21,26]. The ECLS trial similarly searched for eligible participants through primary care EMRs. This trial recruited 12,208 participants and is, consequently, the largest trial for the detection of lung cancer using blood-based biomarkers [9,35]. Additionally, the ECLS and both LHC trials had a lower percentage of respondents drop out between response to invitation and randomisation (see Table 1). This indicates that EMRs can potentially aid researchers in identifying and retaining eligible study subjects.

3. Can Records Be Used to Aid in Identifying Eligible Subjects for Screening?

EMRs have been used to aid in identifying patients eligible for screening. A large-scale study in Minhang District in China, conducted between 2008 and 2016, used EMRs of 5 million patients to identify those eligible for screening multiple cancers including colorectal, gastric, liver, lung, cervical and breast cancer [36]. As a result, more cases of cancer were detected at an early stage, including a number of individuals who were identified as being at high risk of cancer. Similarly, trials for Lung Health Check programmes, implemented in Liverpool and Manchester, were able to recruit and retain a significant proportion of respondents approached [9,26]. These studies indicate that EMRs could be used to conduct more focused interventions. In addition, previous studies have also used machine learning algorithms on smoking history information, identified from EMRs, to create a registry of patients eligible for cancer control efforts, such as smoking cessation and lung cancer screening, which could additionally aid in targeting eligible patients for screening [37,38].

3.1. What Codes Are Associated with LC and Appear in EMRs?

Codes are frequently used to identify patients with various health conditions. Published comorbidity indices and phenotype code lists, such as CALIBER, the Charlson Comorbidity Index, the Elixhauser Comorbidity Index and the Quality and Outcomes Framework (QOF), have compiled a list of codes for lung cancer [39–43]. Moreover, different coding formats are used within different data sources in the EMRs, for example, primary care settings use read codes and secondary care settings use ICD codes [44,45]. A sample code list is presented in Appendix A, Table A1.

Various smoking codes are present within EMRs. These can be used to identify high-risk smokers for screening. Wiley et al. (2013) and Atkinson et al. (2018) examined whether smoking read codes present in EMRs could be used to determine the smoking status of participants [46,47]. Wiley et al. used ICD-9 smoking codes and found that they could accurately detect true smokers in a general population [46]. The combination of codes and free text improved sensitivity to ever smokers, however. Atkinson et al. used smoking read codes found in primary care general practice records to assess participants' smoking history [47]. They found that read codes compared well with a population health survey (Kappa–0.64), indicating that read codes are moderately accurate and, thus, can be used in the identification of smokers.

Codes for health conditions and environmental factors present in EMRs could also be used to identify high-risk groups. A study utilising EMRs from general practices across the UK found that asbestos exposure, COPD and symptoms such as coughing and chest pain were frequently recorded in EMR documentation and prevalent among those diagnosed with lung cancer [48]. Further to this, COPD recording has been explored in EMRs. Algorithms have been developed to determine the presence of COPD in patients. Quint et al. (2014) and Chu et al. (2021) developed two such algorithms that performed well, with positive predictive values (PPVs) of 86.5% and 93.5% [49,50].

Other risk factors such as alcohol consumption and asthma have also been examined. Read codes for alcohol consumption have been validated by comparing EMR data to a health survey. The study by Mansfield et al. (2019) found similar prevalence rates between both a health survey and an EMR dataset, indicating EMRs can be accurately used to identify both current and non-drinkers [51]. Asthma has been validated in EMRs, with the PPVs of studies comparing asthma codes to a reference ranging from 46 to 100% [52].

While there are other social and environmental determinants of lung cancer, such as air pollution and radon exposure, this detailed information is not routinely collected in EMRs. To examine environmental factors, recent studies have linked geospatial and environmental data to EMRs in order to examine related health outcomes [53–55]. Greater consensus on measures to be captured in EMRs, as well as improvements in the linking of external sources of environmental data, could address this issue.

3.2. Use of Free Text to Identify Eligible Participants?

Most studies have used structured variables such as smoking status (non-smoker; ex-smoker; light smoker; moderate smoker; heavy smoker), asthma diagnosed ever (yes/no), pneumonia diagnosed ever (yes/no) and family history of lung cancer (yes/no) to estimate the risk of having lung cancer and to identify participants eligible for lung cancer screening studies [19,56,57]. However, recent studies have begun to explore free text in EMRs to identify eligible patients [58–60].

Natural language processing provides a feasible way to extract various types of information from EMRs. This technique has been successfully used to extract and quantify smoking information in EMRs. De Silva et al. and Palmer et al. used text analysis to quantify pack years from EMR free text [61,62]. This was successfully performed for the majority of cases, but due to the heterogeneity of clinical notes, mis-categorisation and missing cases remained an issue. Smoking status can also be identified accurately from EMRs. Groenhof et al. extracted information on smoking behaviours from free text to categorise participants into current, past and never smokers. Smoking information was accurately retrieved for the majority of cases [63].

This method of smoker identification may be more accurate and less costly and time consuming compared to asking potential participants to fill out questionnaires or to assess their own eligibility for screening. Indeed, free text in EMRs has provided more accurate and comprehensive information on smoking than structured sources of data from EMRs [64]. As these papers indicate that smoking information is present in EMRs and that smokers and non-smokers can be accurately identified from the information contained in them, this method of identification may be feasible for participant identification.

4. What Are the Challenges in Using EMR Data to Detect and Identify High-Risk Populations?

While, when utilising EMR data, screening programmes may achieve better targeting of eligible subjects, there are significant challenges to using EMR data. Data completeness for certain coded data elements can vary, with diagnostic and lifestyle data being less populated than prescription data [62]. Indeed, two prevalent issues affecting data completeness are missing data elements and errors in the recording of health conditions/lifestyle factors. Martin found 43% of the electronic records examined contained errors. Indeed, multiple errors were found in participant records which resulted in a total of 229 errors in 169 participant records [65]. Marston et al.'s study found that 20% of their sample had missing smoking data [66]. While overall trends show that the recording of risk factors such as smoking status has improved, missing data are still a concern, with recorded information on health care indicators only present in 10–40% of sampled EMRs [67–70].

The accuracy and quality of EMR data are a further issue. This is usually examined by comparing coded or extracted EMR data against a "gold standard" reference. Studies examining data quality show mixed results. Booth et al. examined CPRD data compared to population survey data [71]. They found little difference between the prevalence of smoking in CPRD data compared to the population survey. Estimates for current smokers and non-smokers were similar to survey data estimates, but there was underreporting of former smokers in EMRs. Similarly, asthma recording in EMRs was found to compare moderately well with manual chart reviews, with NLP and diagnosis code-based algorithms generating PPVs of 88.0% and 57.1% [72]. Conversely, Modin et al. found significant discordance between pack years recorded in EMRs and pack years determined from a shared decision-making conversation [73]. This research highlights the difficulties in truly determining data accuracy as references may not contain accurate information.

Obtaining ethical approval to access EMRs is equally challenging. EMRs contain sensitive information which means it is imperative that the data are stored and accessed in a secure way. As a result, it can be both costly and time consuming to access and obtain EMR data. Given that the use of EMR data in clinical research has grown, the development and usage of Data Safe Havens to store EMR data have mitigated some of the ethical concerns around the accessibility and storage of the data.

5. Future Research

There has been significant research on the extraction and classification of smoking status in EMRs. However, further research on the use of EMR information to identify and flag patients for follow-ups or screening is required. Safety netting is viewed as a best practice for those at risk of cancer, although there is little evidence for its effectiveness for cancer detection [74]. The use of EMRs to detect and flag patients for follow-ups has been successfully implemented to detect risk of adverse events, delays in follow-ups to abnormal lung imagining findings and delays in cancer diagnosis [75–77]. Algorithms that detect delays in follow-ups have identified a lack of appropriate follow-up action based on four diagnostic cues. The same could be performed to investigate their use for flagging patients that either partially or fully meet screening criteria.

While there is a significant amount of research examining the validity of smoking behaviours in EMRs, further research could be conducted to examine quality for other data elements. There are few papers examining environmental factors such as asbestos and radon exposure. Examining the completeness, accuracy and frequency of recordings for these exposures could aid in identifying high-risk populations.

Further research on lung cancer risk modelling using EMR data is also required [6]. Many risk models have been developed which include clinical and demographic factors. These models utilise trial or registry data and, as a result, there is a lack of research examining the use of real-world EMR information and the use of linked datasets in risk modelling [78]. Wang et al. used EMR data to model the incidence of lung cancer, and they were able to extract a large number of features to include, demonstrating the usefulness of EMR data in modelling [79]. Additionally, further examination of risk models using EMR data would be useful to identify whether models apply well to other datasets.

6. Conclusions

Lung cancer screening using LDCT and biomarkers has the potential to reduce late diagnosis, thereby lowering mortality rates and improving survival of the disease. However, there are significant issues with the detection of subjects eligible for lung cancer screening. Screening trials and programmes have low approach response rates, despite targeting those at a higher risk of developing cancer.

EMRs have provided useful information for clinicians and researchers which has resulted in greater engagement. For example, both the LSUT study and ECLS trial recruited a large number of participants by identifying eligible patients through EMRs. Further, the research presented in this article has shown there are data features contained in EMRs that have the ability to aid screening, such as smoking information contained in codes and free clinical text. This information can ensure that eligible populations are easier to access for researchers/clinicians and that, as a result, these individuals can be better targeted.

There are significant challenges to using EMR data such as a lack of data completeness and data accuracy. With the advances in text analysis and improvements in EMR structure and codes, they may be a viable option that both health systems and researchers can use to identify populations for lung cancer screening.

Author Contributions: Conceptualisation, F.S. and L.B.; introduction, L.B.; Section 1, L.B.; Section 2, U.A.; Section 3, L.B.; future research, L.B.; conclusion, L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Oncimmune Ltd. (grant number SMD0-ZGO002) nand The Melville Trust for the Care and Cure of Cancer (grant number M00109.0001/TZH/MHR).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. Read codes and their associated read terms and conditions.

Read Code	Read Term	Condition
B220100	Malignant neoplasm of mucosa of trachea	Primary Malignancy-Lung
B220.00	Malignant neoplasm of trachea	Primary Malignancy-Lung
B220z00	Malignant neoplasm of trachea NOS	Primary Malignancy-Lung
B221000	Malignant neoplasm of carina of bronchus	Primary Malignancy-Lung
B221100	Malignant neoplasm of hilus of lung	Primary Malignancy-Lung
B221.00	Malignant neoplasm of main bronchus	Primary Malignancy-Lung
B221z00	Malignant neoplasm of main bronchus NOS	Primary Malignancy-Lung
B222000	Malignant neoplasm of upper lobe bronchus	Primary Malignancy-Lung
B222100	Malignant neoplasm of upper lobe of lung	Primary Malignancy-Lung
B222.00	Malignant neoplasm of upper lobe, bronchus or lung	Primary Malignancy-Lung
B222.11	Pancoast’s syndrome	Primary Malignancy-Lung
B222z00	Malignant neoplasm of upper lobe, bronchus or lung NOS	Primary Malignancy-Lung
B223000	Malignant neoplasm of middle lobe bronchus	Primary Malignancy-Lung
B223100	Malignant neoplasm of middle lobe of lung	Primary Malignancy-Lung
B223.00	Malignant neoplasm of middle lobe, bronchus or lung	Primary Malignancy-Lung
B223z00	Malignant neoplasm of middle lobe, bronchus or lung NOS	Primary Malignancy-Lung
B224000	Malignant neoplasm of lower lobe bronchus	Primary Malignancy-Lung
B224100	Malignant neoplasm of lower lobe of lung	Primary Malignancy-Lung
B224.00	Malignant neoplasm of lower lobe, bronchus or lung	Primary Malignancy-Lung
B224z00	Malignant neoplasm of lower lobe, bronchus or lung NOS	Primary Malignancy-Lung
B225.00	Malignant neoplasm of overlapping lesion of bronchus and lung	Primary Malignancy-Lung
B22..00	Malignant neoplasm of trachea, bronchus and lung	Primary Malignancy-Lung
B22y.00	Malignant neoplasm of other sites of bronchus or lung	Primary Malignancy-Lung
B22z.00	Malignant neoplasm of bronchus or lung NOS	Primary Malignancy-Lung

B22z.11	Lung cancer	Primary Malignancy-Lung
BB5S200	[M]Bronchiolo-alveolar adenocarcinoma	Primary Malignancy-Lung
BB5S211	[M]Alveolar cell carcinoma	Primary Malignancy-Lung
BB5S212	[M]Bronchiolar carcinoma	Primary Malignancy-Lung
BB5S400	[M]Alveolar adenocarcinoma	Primary Malignancy-Lung
Byu2000	[X]Malignant neoplasm of bronchus or lung, unspecified	Primary Malignancy-Lung
ZV10100	[V]Personal history of malig neop of trachea/bronchus/lung	Primary Malignancy-Lung
ZV10111	[V]Personal history of malignant neoplasm of bronchus	Primary Malignancy-Lung
ZV10112	[V]Personal history of malignant neoplasm of lung	Primary Malignancy-Lung
ICD10 code	ICD10 term	Condition
C33	Malignant neoplasm of trachea	Primary Malignancy-Lung
C34	Malignant neoplasm of bronchus and lung	Primary Malignancy-Lung

References

1. Cancer Research, U.K. Lung Cancer Statistics. 2021. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer#heading-Zero (accessed on 12 May 2021).
2. Cancer Research, U.K. Lung Cancer Mortality. 2021. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/mortality#heading-Zero (accessed on 12 May 2021).
3. Birring, S.S.; Peake, M.D. Symptoms and the early diagnosis of lung cancer. *Thorax* **2005**, *60*, 268–269.
4. Cancer Research, U.K. Advanced Stage Lung Cancer. 2021. Available online: https://www.cancerresearchuk.org/about-cancer/lung-cancer/advanced/about (accessed on 12 May 2021).
5. Oudkerk, M.; Devaraj, A.; Vliegenthart, R.; Henzler, T.; Prosch, H.; Heussel, C.P.; Bastarrika, G.; Sverzellati, N.; Mascalchi, M.; Delorme, S.; et al. European position statement on lung cancer screening. *Lancet Oncol.* **2017**, *18*, 754–766.
6. De Koning, H.J.; van der Aalst, C.M.; de Jong, P.A.; Scholten, E.T.; Nackaerts, K.; Heuvelmans, M.A.; Lammers, J.W.; Weenink, C.; Yousaf-Khan, U.; Horeweg, N.; et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N. Engl. J. Med.* **2020**, *382*, 503–513.
7. Xu, D.M.; Gietema, H.; de Koning, H.; Vernhout, R.; Nackaerts, K.; Prokop, M.; Weenink, C.; Lammers, J.W.; Groen, H.; Oudkerk, M.; et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. *Lung Cancer* **2006**, *54*, 177–184.
8. Field, J.K.; Duffy, S.W.; Baldwin, D.R.; Brain, K.E.; Devaraj, A.; Eisen, T.; Green, B.A.; Holemans, J.A.; Kavanagh, T.; Kerr, K.M.; et al. The, U.K. Lung Cancer Screening Trial: A pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer. *Health Technol. Assess.* **2016**, *20*, 1–146.
9. Sullivan, F.M.; Mair, F.S.; Anderson, W.; Armory, P.; Briggs, A.; Chew, C.; Dorward, A.; Haughney, J.; Hogarth, F.; Kendrick, D.; et al. Earlier diagnosis of lung cancer in a randomised trial of an autoantibody blood test followed by imaging. *Eur. Respir. J.* **2021**, *57*, 1–11.
10. Wang, G.X.; Baggett, T.P.; Pandharipande, P.V.; Park, E.R.; Percac-Lima, S.; Shepard, J.A.; Fintelmann, F.J.; Flores, E.J. Barriers to lung cancer screening engagement from the patient and provider perspective. *Radiology* **2019**, *290*, 278–287.
11. Lam, S.; Tammemagi, M. Contemporary issues in the implementation of lung cancer screening. *Eur. Respir. Rev.* **2021**, *30*, 1–17.
12. Carter-Harris, L.; Gould, M.K. Multilevel barriers to the successful implementation of lung cancer screening: Why does it have to be so hard? *Ann. Am. Thorac. Soc.* **2017**, *14*, 1261–1265.
13. Thadani, S.R.; Weng, C.; Bigger, J.T.; Ennever, J.F.; Wajngurt, D. Electronic screening improves efficiency in clinical trial recruitment. *J. Am. Med. Inform. Assoc.* **2009**, *16*, 869–873.
14. Malhotra, J.; Malvezzi, M.; Negri, E.; La Vecchia, C.; Boffetta, P. Risk factors for lung cancer worldwide. *Eur. Respir. J.* **2016**, *48*, 889–902.
15. Toumazis, I.; Bastani, M.; Han, S.S.; Plevritis, S.K. Risk-Based lung cancer screening: A systematic review. *Lung Cancer* **2020**, *147*, 154–186.
16. Gandini, S.; Botteri, E.; Iodice, S.; Boniol, M.; Lowenfels, A.B.; Maisonneuve, P.; Boyle, P. Tobacco smoking and cancer: A meta-analysis. *Int. J. Cancer* **2008**, *122*, 155–164.
17. Boyle, P.; Maisonneuve, P. Lung cancer and tobacco smoking. *Lung Cancer* **1995**, *12*, 167–181.

18. Aberle, D.R.; Adams, A.M.; Berg, C.D.; Black, W.C.; Clapp, J.D.; Fagerstrom, R.M.; Gareen, I.F.; Gatsonis, C.; Marcus, P.M.; Sicks, J.D. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **2011**, *365*, 395–409.
19. Katki, H.A.; Kovalchik, S.A.; Berg, C.D.; Cheung, L.C.; Chaturvedi, A.K. Development and validation of risk models to select ever-smokers for CT lung cancer screening. *JAMA* **2016**, *315*, 2300–2311.
20. Ten Haaf, K.; Bastani, M.; Cao, P.; Jeon, J.; Toumazis, I.; Han, S.S.; Plevritis, S.K.; Blom, E.F.; Kong, C.Y.; Tammemägi, M.C.; et al. A comparative modeling analysis of risk-based lung cancer screening strategies. *JNCI J. Natl. Cancer Inst.* **2020**, *112*, 466–479.
21. Rankin, N.M.; McWilliams, A.; Marshall, H.M. Lung cancer screening implementation: Complexities and priorities. *Respirology* **2020**, *25*, 5–23.
22. Paci, E.; Puliti, D.; Pegna, A.L.; Carrozzi, L.; Picozzi, G.; Falaschi, F.; Pistelli, F.; Aquilini, F.; Ocello, C.; Zappa, M.; et al. Mortality, survival and incidence rates in the ITALUNG randomised lung cancer screening trial. *Thorax* **2017**, *72*, 825–831.
23. Becker, N.; Motsch, E.; Trotter, A.; Heussel, C.P.; Dienemann, H.; Schnabel, P.A.; Kauczor, H.U.; Maldonado, S.G.; Miller, A.B.; Kaaks, R.; et al. Lung cancer mortality reduction by LDCT screening—Results from the randomized German, LUSI trial. *Int. J. Cancer* **2020**, *146*, 1503–1513.
24. Quaife, S.L.; Ruparel, M.; Dickson, J.L.; Beeken, R.J.; McEwen, A.; Baldwin, D.R.; Bhowmik, A.; Navani, N.; Sennett, K.; Duffy, S.W.; et al. Lung Screen Uptake Trial (LSUT): Randomized Controlled Clinical Trial Testing Targeted Invitation Materials. *Am. J. Respir. Crit. Care Med.* **2020**, *201*, 965–975.
25. Crosbie, P.A.; Balata, H.; Evison, M.; Atack, M.; Bayliss-Brideaux, V.; Colligan, D.; Duerden, R.; Eaglesfield, J.; Edwards, T.; Elton, P.; et al. Implementing lung cancer screening: Baseline results from a community-based 'Lung Health Check'pilot in deprived areas of Manchester. *Thorax* **2019**, *74*, 405–409.
26. Ghimire, B.; Maroni, R.; Vulkan, D.; Shah, Z.; Gaynor, E.; Timoney, M.; Jones, L.; Arvanitis, R.; Ledson, M.; Lukehirst, L.; et al. Evaluation of a health service adopting proactive approach to reduce high risk of lung cancer: The Liverpool Healthy Lung Programme. *Lung Cancer* **2019**, *134*, 66–71.
27. Ali, N.; Lifford, K.J.; Carter, B.; McRonald, F.; Yadegarfar, G.; Baldwin, D.R.; Weller, D.; Hansell, D.M.; Duffy, S.W.; Field, J.K.; et al. Barriers to uptake among high-risk individuals declining participation in lung cancer screening: A mixed methods analysis of the UK Lung Cancer Screening (UKLS) trial. *BMJ Open* **2015**, *5*, e008254.
28. Patel, D.; Akporobaro, A.; Chinyanganya, N.; Hackshaw, A.; Seale, C.; Spiro, S.G.; Griffiths, C. Attitudes to participation in a lung cancer screening trial: A qualitative study. *Thorax* **2012**, *67*, 418–425.
29. Quaife, S.L.; Marlow, L.A.; McEwen, A.; Janes, S.M.; Wardle, J. Attitudes towards lung cancer screening in socioeconomically deprived and heavy smoking communities: Informing screening communication. *Health Expect.* **2017**, *20*, 563–573.
30. Chapple, A.; Ziebland, S.; McPherson, A. Stigma, shame, and blame experienced by patients with lung cancer: Qualitative study. *BMJ* **2004**, *328*, 1470.
31. Van Hal, G.; Garcia, P.D. Lung cancer screening: Targeting the hard to reach—a review. *Transl. Lung Cancer Res.* **2021**, *10*, 2309–2322.
32. Williamson, T.J.; Rawl, S.M.; Kale, M.S.; Carter-Harris, L. Lung cancer screening and stigma: Do smoking-related differences in perceived lung cancer stigma emerge prior to diagnosis? *Stigma Health* **2021**, https://doi.org/10.1093/abm/kaz063.
33. Carter-Harris, L. Lung cancer stigma as a barrier to medical help-seeking behavior: Practice implications. *J. Am. Assoc. Nurse Pract.* **2015**, *27*, 240–245.
34. Vrinten, C.; Gallagher, A.; Waller, J.; Marlow, L.A. Cancer stigma and cancer screening attendance: A population based survey in England. *BMC Cancer* **2019**, *19*, 566.
35. Sullivan, F.M.; Farmer, E.; Mair, F.S.; Treweek, S.; Kendrick, D.; Jackson, C.; Robertson, C.; Briggs, A.; McCowan, C.; Bedford, L.; et al. Detection in blood of autoantibodies to tumour antigens as a case-finding method in lung cancer using the EarlyCDT®-Lung Test (ECLS): Study protocol for a randomized controlled trial. *BMC Cancer* **2017**, *17*, 1.
36. He, D.; Xu, W.; Su, H.; Li, W.; Zhou, J.; Yao, B.; Xu, D.; He, N. Electronic health Record-Based screening for major cancers: A 9-year experience in Minhang district of Shanghai, China. *Front. Oncol.* **2019**, *9*, 375.
37. Onega, T.; Nutter, E.L.; Sargent, J.; Doherty, J.A.; Hassanpour, S. Identifying patient smoking history for cessation and lung cancer screening through mining electronic health records. *Cancer Epidemiol. Prev. Biomark.* **2017**, *26*, 437.
38. Hippisley-Cox, J.; Coupland, C. Identifying patients with suspected lung cancer in primary care: Derivation and validation of an algorithm. *Br. J. Gen. Pract.* **2011**, *61*, 715–723.
39. Kuan, V.; Denaxas, S.; Gonzalez-Izquierdo, A.; Direk, K.; Bhatti, O.; Husain, S.; Sutaria, S.; Hingorani, M.; Nitsch, D.; Parisinos, C.A.; et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit. Health* **2019**, *1*, 63–77.
40. Charlson, M.E.; Pompei, P.; Ales, K.L.; MacKenzie, C.R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **1987**, *40*, 373–383.
41. Elixhauser, A.; Steiner, C.; Harris, D.R.; Coffey, R.M. Comorbidity measures for use with administrative data. *Med Care* **1998**, *36*, 8–27.
42. Metcalfe, D.; Masters, J.; Delmestri, A.; Judge, A.; Perry, D.; Zogg, C.; Gabbe, B.; Costa, M. Coding algorithms for defining Charlson and Elixhauser co-morbidities in Read-coded databases. *BMC Med. Res. Methodol.* **2019**, *19*, 115.
43. NHS Digital, Quality and Outcomes Framework. 2020. Available online: https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data (accessed on 21 May 2021).

44. NHS Digital, Read Codes. 2020. Available online: https://digital.nhs.uk/services/terminology-and-classifications/read-codes (accessed on 21 May 2021).
45. NHS Digital, NHS classification service. *NHS Digit. Trud.* **2021**. Available online: https://isd.digital.nhs.uk/trud3/user/guest/group/0/home (accessed on 20 May 2021).
46. Wiley, L.K.; Shah, A.; Xu, H.; Bush, W.S. ICD-9 tobacco use codes are effective identifiers of smoking status. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 652–658.
47. Atkinson, M.D.; Kennedy, J.I.; John, A.; Lewis, K.E.; Lyons, R.A.; Brophy, S.T. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 2.
48. Soriano, L.C.; Zong, J.; Rodríguez, L.A. Feasibility and validity of The Health Improvement Network database of primary care electronic health records to identify and characterise patients with small cell lung cancer in the United Kingdom. *BMC Cancer* **2019**, *19*, 1–9.
49. Quint, J.K.; Müllerova, H.; DiSantostefano, R.L.; Forbes, H.; Eaton, S.; Hurst, J.R.; Davis, K.; Smeeth, L. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open* **2014**, *4*, e005540.
50. Chu, S.H.; Wan, E.S.; Cho, M.H.; Goryachev, S.; Gainer, V.; Linneman, J.; Scotty, E.J.; Hebbring, S.J.; Murphy, S.; Lasky-Su, J.; et al. An independently validated, portable algorithm for the rapid identification of COPD patients using electronic health records. *Sci. Rep.* **2021**, *11*, 1–9.
51. Mansfield, K.; Crellin, E.; Denholm, R.; Quint, J.K.; Smeeth, L.; Cook, S.; Herrett, E. Completeness and validity of alcohol recording in general practice within the UK: A cross-sectional study. *BMJ Open* **2019**, *9*, http://dx.doi.org/10.1136/bmjopen-2019-031537.
52. Nissen, F.; Quint, J.K.; Wilkinson, S.; Mullerova, H.; Smeeth, L.; Douglas, I.J. Validation of asthma recording in electronic health records: A systematic review. *Clin. Epidemiol.* **2017**, *9*, 643.
53. Schinasi, L.H.; Auchincloss, A.H.; Forrest, C.B.; Roux, A.V. Using electronic health record data for environmental and place based population health research: A systematic review. *Ann. Epidemiol.* **2018**, *28*, 493–502.
54. Torres-Durán, M.; Casal-Mouriño, A.; Ruano-Ravina, A.; Provencio, M.; Parente-Lamelas, I.; Hernández-Hernández, J.; Vidal-García, I.; Varela-Lema, L.; Valdés Cuadrado, L.; Fernández-Villar, A.; et al. Residential radon and lung cancer characteristics at diagnosis. *Int. J. Radiat. Biol.* **2021**, https://doi.org/10.1080/09553002.2021.1913527.
55. Boulos, M.N. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. *Int. J. Health Geogr.* **2004**, *3*, 1–50.
56. Okoli, G.N.; Kostopoulou, O.; Delaney, B.C. Is symptom-based diagnosis of lung cancer possible? A systematic review and meta-analysis of symptomatic lung cancer prior to diagnosis for comparison with real-time data from routine general practice. *PLoS ONE* **2018**, *13*, e0207686.
57. Klingman, K.J.; Sprey, J. Insomnia disorder diagnosis and treatment patterns in primary care: A cross-sectional analysis of electronic medical records data. *J. Am. Assoc. Nurse Pract.* **2020**, *32*, 145–151.
58. Solarte-Pabon, O.; Torrente, M.; Rodriguez-González, A.; Provencio, M.; Menasalvas, E.; Tuñas, J.M. Lung cancer diagnosis extraction from clinical notes written in spanish. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 22–24 June, 28–30 July 2020; pp. 492–497.
59. Ruiz, E.M.; Tuñas, J.M.; Bermejo, G.; Martín, C.G.; Rodríguez-González, A.; Zanin, M.; de Pedro, C.G.; Méndez, M.; Zaretskaia, O.; Rey, J.; et al. Profiling lung cancer patients using electronic health records. *J. Med. Syst.* **2018**, *42*, 1–10.
60. Jensen, K.; Soguero-Ruiz, C.; Mikalsen, K.O.; Lindsetmo, R.O.; Kouskoumvekaki, I.; Girolami, M.; Skrovseth, S.O.; Augestad, K.M. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci. Rep.* **2017**, *7*, 1–2.
61. Palmer, E.L.; Hassanpour, S.; Higgins, J.; Doherty, J.A.; Onega, T. Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1–10.
62. De Silva, L.; Ginter, T.; Forbush, T.; Nokes, N.; Fay, B.; Mikuls, T.; Cannon, G.; DuVall, S. Extraction and quantification of pack-years and classification of smoker information in semi-structured Medical Records. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June 2011; pp. 1–8.
63. Groenhof, T.K.; Koers, L.R.; Blasse, E.; de Groot, M.; Grobbee, D.E.; Bots, M.L.; Asselbergs, F.W.; Lely, A.T.; Haitjema, S.; van Solinge, W.; et al. Data mining information from electronic health records produced high yield and accuracy for current smoking status. *J. Clin. Epidemiol.* **2020**, *118*, 100–106.
64. Wang, L.; Ruan, X.; Yang, P.; Liu, H. Comparison of three information sources for smoking information in electronic health records. *Cancer Inform.* **2016**, *15*, 237–242.
65. Martin, P.M. Can we trust electronic health records? The smoking test for commission errors. *BMJ Health Care Inform.* **2018**, *25*, http://dx.doi.org/10.14236/jhi.v25i2.970.
66. Marston, L.; Carpenter, J.R.; Walters, K.R.; Morris, R.W.; Nazareth, I.; White, I.R.; Petersen, I. Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: A cross-sectional study. *BMJ Open* **2014**, *4*, 1–7.
67. Simpson, C.R.; Hippisley-Cox, J.; Sheikh, A. Trends in the epidemiology of smoking recorded in UK general practice. *Br. J. Gen. Pract.* **2010**, *60*, 121–127.
68. Szatkowski, L.; Lewis, S.; McNeill, A.; Coleman, T. Is smoking status routinely recorded when patients register with a new GP? *Fam. Pract.* **2010**, *27*, 673–675.

69. Thiru, K.; Hassey, A.; Sullivan, F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ* **2003**, *326*, 1070.
70. Petersen, I.; Welch, C.A.; Nazareth, I.; Walters, K.; Marston, L.; Morris, R.W.; Carpenter, J.R.; Morris, T.P.; Pham, T.M. Health indicator recording in UK primary care electronic health records: Key implications for handling missing data. *Clin. Epidemiol.* **2019**, *11*, 157–167.
71. Booth, H.P.; Prevost, A.T.; Gulliford, M.C. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011. *Pharmacoepidemiol. Drug Saf.* **2013**, *22*, 1357–1361.
72. Wu, S.T.; Sohn, S.; Ravikumar, K.E.; Wagholikar, K.; Jonnalagadda, S.R.; Liu, H.; Juhn, Y.J. Automated chart review for asthma cohort identification using natural language processing: An exploratory study. *Ann. Allergy Asthma Immunol.* **2013**, *111*, 364–369.
73. Modin, H.E.; Fathi, J.T.; Gilbert, C.R.; Wilshire, C.L.; Wilson, A.K.; Aye, R.W.; Farivar, A.S.; Louie, B.E.; Vallières, E.; Gorden, J.A. Pack-year cigarette smoking history for determination of lung cancer screening eligibility. Comparison of the electronic medical record versus a shared decision-making conversation. *Ann. Am. Thorac. Soc.* **2017**, *14*, 1320–1325.
74. Nicholson, B.D.; Mant, D.; Bankhead, C. Can safety-netting improve cancer detection in patients with vague symptoms? *BMJ* **2016**, *355*, https://doi.org/10.1136/bmj.i5515.
75. Murphy, D.R.; Laxmisan, A.; Reis, B.A.; Thomas, E.J.; Esquivel, A.; Forjuoh, S.N.; Parikh, R.; Khan, M.M.; Singh, H. Electronic health record-based triggers to detect potential delays in cancer diagnosis. *BMJ Qual. Saf.* **2014**, *23*, 8–16.
76. Murphy, D.R.; Thomas, E.J.; Meyer, A.N.; Singh, H. Development and validation of electronic health record–based triggers to detect delays in follow-up of abnormal lung imaging findings. *Radiology* **2015**, *277*, 81–87.
77. Hinrichsen, V.L.; Kruskal, B.; O'Brien, M.A.; Lieu, T.A.; Platt, R. Using electronic medical records to enhance detection and reporting of vaccine adverse events. *J. Am. Med. Inform. Assoc.* **2007**, *14*, 731–735.
78. Department of Health and Social Care, Data saves lives: Reshaping health an social care with data (draft). 2021. Available online: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data-draft/data-saves-lives-reshaping-health-and-social-care-with-data-draft (accessed on 10 August 2021).
79. Wang, X.; Zhang, Y.; Hao, S.; Zheng, L.; Liao, J.; Ye, C.; Xia, M.; Wang, O.; Liu, M.; Weng, C.H.; et al. Prediction of the 1-year risk of incident lung cancer: Prospective study using electronic health records from the state of maine. *J. Med. Internet Res.* **2019**, *21*, 1–17.