

Developing a risk prediction tool for Lung Cancer in Kent and Medway, England: Cohort Study using linked Data

David Howell (✉ david@quantum-analytica.co.uk)

Quantum Analytica

Ross Buttery

Quantum Analytica

Padmanabhan Badrinath

Kent County Council

Abraham George

Kent County Council

Rithvik Hariprasad

Ian Vousden

NHS England Kent Cancer Alliance

Tina George

NHS England Kent Cancer Alliance

Cathy Finnis



NHS England Kent Cancer Alliance

Article

Keywords:

Posted Date: June 28th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3100044/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Lung cancer has the poorest survival due to late diagnosis and there is no universal screening. Hence early detection is crucial. Our objective was to develop a lung cancer risk prediction tool at a population level.

Methods

We used a large place based linked data set from a local health system in southeast England which contained extensive information on each individual covering demographic, socioeconomic, lifestyle, health, and care service utilization. We exploited the power of Machine Learning to derive risk scores using linear regression modelling. Tens of thousands of model runs were undertaken to identify attributes which predicted the risk of lung cancer.

Results

Initially sixteen attributes were identified. A final combination of seven attributes were chosen based on the number of cancers detected which formed the Kent & Medway lung cancer risk prediction tool. This was then compared with the criteria used in the wider Targeted Lung Health Checks programme. The prediction tool outperformed by detecting 822 cases compared to 581 by the lung check programme currently in operation.

Conclusion

We have demonstrated the exceptional application and utility of Machine Learning in developing a risk score for lung cancer and discuss its clinical applicability.

Introduction

Lung cancer is one of the major causes of death worldwide^{1,2} and in the UK around 48,500 new lung cancer cases are detected every year and it is the third most common cancer. Every day 95 people die of lung cancer with an annual total of 34,800³. Smoking is a major risk factor and 90% of the world's cases are caused by cigarette consumption⁴. Furthermore, there is an association with prolonged environmental exposure to air pollutants such as sulphur, nitrogen, or arsenic; hence, nations with greater levels of pollution are likely to have higher incidences of lung cancer⁵. Until the advent of the Targeted Lung Health Check (TLHC) pilots, it was only when a person started to exhibit the symptoms of lung cancer, that a diagnosis of the disease could be made. Some of these symptoms could include coughing, shortness of breath, unexplained weight loss, wheezing, haemoptysis, chest discomfort, exhaustion, and decreased appetite⁶.

The lack of overt symptoms in the early stages of lung cancer often leads to patients presenting late resulting in delayed diagnosis and treatment. The escalating fatality rate can be attributed to patients seeking medical

attention at advanced stages of the disease, diminishing the prospects of successful surgical removal and intervention⁷. A study⁸ published in the British Journal of Cancer examined the relationship between stage at diagnosis, early mortality and major demographic variables. The authors found that around 70% of cases of lung cancer are diagnosed at a late stage, after it has metastasized and spread into other parts of the body. Late diagnosis results in a far greater mortality and early diagnosis can therefore make a significant difference in outcome for the patient. According to Cancer Research UK the “the proportion of people surviving their cancer for five years or more is around 6 in 10 if diagnosed at earlier stage and less than 1 in 10 if diagnosed at the latest stage”⁹. Not only would this have markedly improved live expectancy, but patient morbidity is also significantly improved as a result of larger functioning lung capacity following tumor removal. The National Health Service (NHS) Long Term Plan¹⁰ sets out the ambition that by 2028 the proportion of cancer diagnosed at an early stage will rise from around half, which is the current position taking all cancers together, to three-quarters of cancer patients”.

Treatment of lung cancer is considerable and varies depending on the stage of diagnosis. If cancer is detected at the early operable stage of lung cancer, then the primary treatment costs predominantly involve surgical removal procedures¹¹. However, as the disease advances to Stage 3 and Stage 4, the expenses associated with surgical interventions tend to decrease, while the costs related to chemotherapy escalate significantly. This shift in treatment modalities is primarily due to the diminished feasibility of surgical removal as the cancer spreads and becomes more widespread. Instead, chemotherapy becomes a pivotal component of the treatment regimen during the advanced stages, aiming to control tumor growth, alleviate symptoms, and potentially prolong survival. Consequently, the timely identification and detection of lung cancer can significantly alleviate the financial burden on the state, the insurer or the patients and their families. This includes mitigating the expenses associated with advanced-stage treatments, extended hospital stays, intensive therapies, and palliative care services. Moreover, early detection may allow for a wider range of treatment options, including less invasive procedures, targeted therapies, and improved chances of successful outcomes.

Artificial Intelligence (AI) is a new and rapidly evolving field where computers are taught to think like humans. Due to its enhanced accuracy, precision, and decision support capabilities, AI has begun to be implemented in modern medicine. It is being used in two ways namely, physical and virtual. Physical applications of AI include robots that are automated to perform tasks such as caring for the elderly and others that assist in surgeries. ML is a subfield of AI that deals with the virtual aspect. ML models can be trained to detect or predict occurrences of a health condition¹². AI is suitable in the medical field as it has no concept of fatigue unlike doctors and therefore can process large number of images and data at any given time¹³. This requires a good prediction model to be designed which involves acquiring a large dataset for training the model. The bigger and more diverse the dataset is, better the results that can be expected from it¹⁴.

With the help of AI we can make accurate assessments of one’s risk of lung cancer. The detection or prediction of lung cancer serves as a prime illustration where the utilisation of AI is indispensable. This is due to the fact that lung cancer is a highly time sensitive condition and early diagnosis can be difference between life and death. Risk factors associated with lifestyle choices can be used to provide profiles of potential risks for each

person. This may provide a precise means to determine individuals who are more prone to lung cancer and thereby raise awareness for earlier detection and treatment¹⁵.

Currently there are no accepted screening methods for lung cancer that provide socio-economic benefits to the healthcare system^{16,17}. Current attempts made to improve early lung cancer diagnosis involve diagnostically evaluating large volumes of individuals with less than 1% of successful case identification^{18,19}. The population of England is estimated to increase by 6% over the next decade²⁰. Furthermore, there has been a 19% increase in the prevalence of cancer in England over the last decade and published figures on the number of people waiting for a diagnosis or treatment for cancer have shown the huge challenge facing NHS cancer services, with tens of thousands of people waiting too long for diagnosis or vital treatment, especially since the start of the pandemic of COVID-19²¹. Hence the NHS cannot afford to provide existing healthcare in the same way in the future and will not have a sufficient workforce to deliver this. This challenge is not just isolated to the UK but is a common issue worldwide.

Our study aims to address the challenge of delayed diagnosis of lung cancer by exploiting the processing power of AI. We developed a model for providing risk-based predictions of lung cancer based on an individual's lifestyle choices, family history and other clinical data. We had access to a large dataset consisting of 1.25 million adult residents across the Kent and Medway region called the Kent Integrated Dataset (KID)²². We harnessed the capabilities of ML to train the model in making risk predictions by extracting patterns from data records of residents who had been diagnosed and treated for lung cancer. Our objective was to find the best performing model among a group of ML algorithms that gave accurate predictions of the risk of lung cancer.

Methods

The County of Kent:

Kent is the largest county in England with a population of 1.6 million²³. It has an exceptional spread of affluence and extreme poverty. Before COVID, a life expectancy gap of almost 20 years already existed between the least and most deprived wards²⁴. Some of the largest groups which suffer extreme health inequalities are asylum seekers, migrants and refugees, Gypsy, Roma and Travelers, veterans, looked after children and seasonal agricultural workers. Kent is faced with a range of key health challenges. Widening inequalities in health and wellbeing are observed across both geographical areas and amongst people with different vulnerabilities influenced by a range of wider determinants of health. A 'coastal excess or effect' in health inequalities exists across its numerous coastal and rural communities²⁵.

Dataset description:

Data for this study was taken from the KID²², which contains a vast array of patient level, pseudonymised integrated health and care data. The KID is overseen by a steering group known as the Kent & Medway Shared Health and Care Analytics board (SHcAB) that includes representatives of Kent County Council, local health commissioners and information governance leads. The SHcAB considers issues such as information governance, development of the dataset and applications for use of the data. The Kent and Medway data

warehouse team provides day-to-day administration and project management. Access was granted to the first author by the SHcAB for the study duration through established due process. Patients can opt-out of contributing data to the KID by informing their GP surgery that they do not want their data to be shared with external organisations. It has to be appreciated that the data is not in the public domain as it is a pseudonymised person level data set. We established a project oversight group, supported by the Kent & Medway cancer alliance which included cancer clinicians, service managers, Public Health physicians, epidemiologists, and AI experts. Regular stakeholder engagement took place throughout the study involving patients and public representatives.

Data contained within the KID represented a six-year longitudinal record of health and care data for residents for 2014-2019 which was 1,865,382. An initial exclusion for under 18s years was made (n=599,866) which reduced the cohort to 1,265,516. We then removed a further 10,532 patients (0.8% of the total population), due to incomplete or missing records data, which took the original cohort size to 1,254,984. The final dataset contained a total of 1,254,984 patients of which 6053 were diagnosed with a primary lung cancer during this period and these were included within the scope of this investigation. The cohort selection (lung cancer cohort) only encompassed lung cancers that originated from a primary metastatic tumor site, effectively excluding benign tumors and secondary metastases caused by other types of cancer. To ensure comprehensive capture of all patients meeting the criteria, we assessed both primary and secondary healthcare records using relevant SNOMED or ICD-10 codes respectively. Patients with Lung Cancer included all confirmed diagnoses regardless of diagnosis of care setting, staging at the time of diagnosis, disease progression or onward treatment options and outcomes. Core dimensions of data used within this study are shown below:

- Patient Demographics
- Primary Care (Events, Consultations, Long term condition registers, Medications, Deaths)
- Secondary Care (A&E, Inpatient Spells and Outpatients, Critical Care Bed Days)
- Mental Health (Inpatient and Outpatient History)
- Community Care (Contacts, Appointments, Minor Injuries Units and Walk In Centers)
- Wider Determinants of Health including Housing, Education, Occupation, Economic and Deprivation
- Environmental Datasets - Pollution, Radon ground levels

Data Pre-processing:

The dataset contained missing values mainly in the attribute named 'ethnicity' as shown in Table 1, despite a lot of work to try and capture ethnicity coding from various sources. We therefore excluded this from the model as we felt that it wasn't appropriate to try and use average value or synthetic data derivative which is done in most cases. Other dataset attributes had limited to no missing or outlier values from features, so no further transformations were made on the remainder of the datasets.

The data attributes are grouped into life history, symptoms, diagnostics, treatment and end of life care based on the stage at which the data is collected as depicted in figure 1. To prepare the model for predicting patients' risk ratios, we extracted only the essential attributes from the dataset. These columns were selected based on their potential to provide valuable predictive information. We specifically focused on data concerning the pathways leading to the diagnosis of lung cancer as it held valuable insights regarding the associated causes and symptoms. Attributes related to cancer diagnosis or data related to two-week wait urgent referrals, appointments to see an oncologist, Chest X-Rays and Low Dose Computer Aided Tomography (LDCT) scans for confirming diagnosis, treatment options such as chemotherapy and radiotherapy and mortality were omitted. These attributes were excluded from the dataset because they were deemed as non-predictive elements that did not offer significant insights into the associated risks of a positive diagnosis of lung cancer. We excluded the above diagnostics and treatment elements up to 12 months before the date of diagnosis.

Relative risks (RR) were calculated for all the variables and were used to determine the important attributes and for categorisation. Relative risk is the ratio of the incidence of an event occurring (Lung Cancer) with an exposure (e.g., smoking) versus the incidence of the same event occurring without the exposure. For example, the relative risk of developing lung cancer in smokers (the exposed group) versus non-smokers (non-exposed group) would be the probability of developing lung cancer for smokers divided by the probability of developing lung cancer for non-smokers. All characteristics of the individual datasets such as medications, events, tests, demographic qualities or wider determinant of health factors were tested, and risk scored using this methodology. To reduce the number of categories we collapsed these into meaningful groupings and these were informed by the higher relative risk of related variables. For instance, for respiratory disorders such as COPD and Asthma each of which have numerous diagnosis codes, these were built up into simple three state options; Yes, No or Has Familial History. Other features such as smoking history and activity with high dimensionality were ranked into similar groups by creating scores.

Model development:

We used feature encoding to reduce the number of states and to simplify the complexity of model development and enhance performance. One-hot encoding and standard scaling was used for the feature encoding²⁶. Given the need to develop a scalar response to risk scoring in order to aid prioritisation of patients at greatest risk of developing lung cancer within a screening pool, logistical and other categorical models were ruled out. Traditional linear regression was selected as an initial candidate model to detect lung cancers early and thereby improving outcomes over and above the current screening protocol for lung cancer in the UK.

Using a combination of methods namely informed by the data, proposals from clinical experts and published literature^{27,28}, sixteen attributes were identified. We took our entire population data for n attributes, which could be anywhere between 2 to 16, and split this into 70% training and 30% validation datasets²⁹. We then used the 70% dataset to build a linear regression model on these n attributes. We developed a loop within Python³⁰ to identify all the possible combinations of these 16 attributes in their ability to detect lung cancer. We applied this model for n attributes to the 30% test population to achieve an output which is number of lung cancer cases detected. This was repeated one hundred times (Figure 2) in order to create multiple outputs that could be averaged to test the models' repeatability and for onward evaluation. We then employed boot strapping³¹ to test the general ability of the model to work across randomised populations. In each run, both

the 70% training set and the 30% validation set were again randomized to eliminate any potential biases or chance influences. This randomization also aimed to provide comprehensive average performance statistics for all models. In each model run the TLHC eligibility criteria were applied, and the number of cancers counted. This was compared to the highest risk scored patients identified by the prediction model, keeping both the screening cohort sizes equal.

Model evaluation: Evaluation of the algorithm could not be investigated using standardised evaluation methods (e.g., R^2) due to the desired scalar output of the model as our objective was to identify a cohort at high risk of lung cancer. Instead, we rationalised that if the algorithm is working most efficiently, we should be able to demonstrate more lung cancer cases being found within a screening pool in the population compared to that of the current screening pilots ongoing in England. In order to baseline our evaluation therefore, we compared the output of the algorithm against the current screening population for the TLHC³² programme. Patients meeting the following three criteria will be invited for screening:

- are over 55 but younger than 75 years old
- are registered with an GP in the area the scheme is operating
- have ever smoked, and this is recorded with the GP.

This number of cases found from the TLHC programme was then compared with the number of cases identified using the linear regression model using the top performing combination of attributes.

Results

Selected characteristics of cohorts included in the study are shown in Table 1.

Table 1: Few selected characteristics of cohorts included in the study.

Features	Lung Cancer Cohort (n= 6053)	Non-lung cancer cohort (n= 1248931)	Whole cohort (n= 1254984)
Age (Years)			
18-25	103 (1.8%)	150304 (12%)	150407 (12%)
26-44	642 (10.6%)	378802 (30.5%)	379444 (30.3%)
45-59	1241 (20.5%)	324581 (26%)	325822 (26%)
60+	4067 (67.1%)	395244 (31.5%)	399311 (31.7%)
Gastroenterological Disorders			
Yes	537 (8.9%)	55814 (4.5%)	56351 (4.5%)
No	5516 (91.1%)	1193117 (95.5%)	1198633 (95.5%)
Race (%)			
White – British	2281 (37.8%)	411159 (33.1%)	413440 (33.1%)
White - Any other White background	62 (1%)	17838 (1.4%)	17900 (1.4%)
White – Irish	18 (0.4%)	1626 (0.1%)	1644 (0.1%)
Black or Black British - Caribbean	2 (0%)	813 (0.1%)	815 (0.1%)
Black or Black British – African	11 (0.2%)	3667 (0.3%)	3678 (0.3%)
Black or Black British - Any other Black background	6 (0.1%)	2002 (0.2%)	2008 (0.2%)
Asian or Asian British - Bangladeshi	1 (0%)	832 (0.1%)	833 (0.1%)
Asian or Asian British - Pakistani	2 (0%)	746 (0.1%)	748 (0.1%)
Asian or Asian British – Indian	15 (0.2%)	5640 (0.5%)	5655 (0.5%)
Asian or Asian British - Any other Asian background	12 (0.2%)	3762 (0.3%)	3774 (0.3%)
Mixed - White and Black African	0 (0%)	485 (0%)	485 (0%)
Mixed - White and Black Caribbean	0 (0%)	591 (0%)	591 (0%)

Mixed - White and Asian	1 (0%)	762 (0.1%)	763 (0.1%)
Mixed - Any other mixed background	2 (0%)	1868 (0.1%)	1870 (0.1%)
Other Ethnic Groups – Chinese	3 (0%)	928 (0.1%)	931 (0.1%)
Other Ethnic Groups - Any other ethnic group	19 (0.3%)	5344 (0.4%)	5363 (0.4%)
Not stated	680 (11.3%)	98172 (7.9%)	98852 (7.9%)
Not known	2938 (48.5%)	686643 (55.2%)	689581 (55.2%)
Smoking Status (%)			
Never Smoked	968 (16%)	392289 (31.4%)	393257 (31.4%)
Passive Smoker / Ex-Trivial Smoker (<1 a day)	1110 (18.3%)	275656 (22.1%)	276766 (22.1%)
Trivial Smoker (<1 a day) / Ex-Light Smoker (1 - 9 a day)	691 (11.4%)	141641 (11.3%)	142332 (11.3%)
Light Smoker (1 - 9 a day) / Ex-Moderate Smoker (10 - 19 a day)	1117 (18.5%)	222730 (17.8%)	223847 (17.8%)
Moderate Smoker (10 - 19 a day) / Ex-Heavy Smoker (20+ a day)	1745 (28.8%)	186827 (15%)	188572 (15%)
Heavy Smoker (20+ a day)	422 (7%)	29788 (2.4%)	30210 (2.4%)
Care Home (%)			
Care Home	51 (0.8%)	6946 (0.6%)	6997 (0.6%)
Not in a Care Home	6002 (99.2%)	1241985 (99.4%)	1247987 (99.4%)
Deprivation (Decile)			
1 - Most Deprived	390 (6.4%)	75207 (6%)	75597 (6%)
2	546 (9%)	107944 (8.6%)	108490 (8.6%)
3	525 (8.7%)	105137 (8.4%)	105662 (8.4%)
4	665 (11%)	126404 (10.1%)	127069 (10.1%)
5	812 (13.4%)	158157 (12.7%)	158969 (12.7%)
6	683 (11.3%)	134822 (10.8%)	135505

			(10.8%)
7	776 (12.8%)	165452 (13.2%)	166228 (13.2%)
8	655 (10.8%)	127382 (10.2%)	128037 (10.2%)
9	516 (8.5%)	117639 (9.4%)	118155 (9.4%)
10 - Least Deprived	451 (7.5%)	119876 (9.6%)	120327 (9.6%)
Unknown	34 (0.6%)	10911 (0.9%)	10945 (0.9%)
Population Segmentation Clusters (ACORN)			
Affluent Achievers	1490 (24.6%)	297983 (24%)	299473 (24%)
Comfortable Communities	1905 (31.5%)	381269 (31%)	383174 (31%)
Financially Stretched	1364 (22.5%)	256201 (21%)	257565 (21%)
Not Private Households	45 (0.7%)	8563 (1%)	8608 (1%)
Rising Prosperity	233 (3.8%)	68672 (6%)	68905 (6%)
Urban Adversity	707 (11.7%)	164400 (13%)	165107 (13%)
Undefined	309 (5.2%)	71843 (6%)	72152 (6%)
COPD			
Yes	1579 (26.1%)	185039 (14.8%)	186618 (14.8%)
No	4306 (71.1%)	1020885 (81.8%)	1025191 (81.8%)
Family History	168 (2.8%)	43007 (3.4%)	43175 (3.4%)
Hypertension			
Yes	1855 (30.6%)	210788 (16.9%)	212643 (16.9%)
No	3900 (64.4%)	952750 (76.3%)	956650 (76.3%)
Family History	298 (5%)	85393 (6.8%)	85691 (6.8%)
Diabetes			

Yes	2003 (33.1%)	278378 (22.2%)	280381 (22.2%)
No	3953 (65.3%)	943729 (75.6%)	947682 (75.6%)
Family History	97 (1.6%)	26824 (2.2%)	26921 (2.2%)
Tuberculosis			
Yes	75 (1.2%)	4823 (0.4%)	4898 (0.4%)
No	5961 (98.5%)	1242324 (99.5%)	1248285 (99.5%)
Family History	17 (0.3%)	1784 (0.1%)	1801 (0.1%)
Activity (%)			
Competitive Athlete	1 (0%)	267 (0%)	268 (0%)
Heavy (3+ days a week)	342 (5.7%)	90414 (7.2%)	90756 (7.2%)
Intermediate (2 Days a week)	4092 (67.6%)	905749 (72.5%)	909841 (72.5%)
Light (1 day a week)	912 (15%)	143704 (11.6%)	144616 (11.6%)
Rarely (<1 day a week)	652 (10.8%)	103798 (8.3%)	104450 (8.3%)
Exercise Impossible	54 (0.9%)	4999 (0.4%)	5053 (0.4%)
Other Cancers			
Yes (excludes lung cancer)	1281 (21.2%)	116998 (9.4%)	118279 (9.4%)
No	4354 (71.9%)	1058046 (84.7%)	1062400 (84.7%)
Family History	418 (6.9%)	73887 (5.9%)	74305 (5.9%)
Cardiac Disorders			
Yes	2093 (34.6%)	207638 (16.6%)	209731 (16.7%)
No	3436 (56.8%)	991171 (79.4%)	994607 (79.3%)
Family History	524 (8.7%)	50122 (4%)	50646 (4%)
Respiratory Disorders			

Yes	3845 (63.5%)	670351 (53.7%)	674196 (53.7%)
No	2122 (35.1%)	559762 (44.8%)	561884 (44.8%)
Family History	86 (1.4%)	18818 (1.5%)	18904 (1.5%)
Male			
Yes	2916 (48.2%)	607295 (48.6%)	610211 (48.6%)
No	3137 (51.8%)	641627 (51.4%)	644764 (51.4%)
Unknown	0 (0%)	9 (0%)	9 (0%)
Female			
Yes	3137 (51.8%)	641627 (51.4%)	644764 (51.4%)
No	2916 (48.2%)	607295 (48.6%)	610211 (48.6%)
Family History	0 (0%)	9 (0%)	9 (0%)

Relative risks for the attributes included in the model are presented in Table 2.

In the attribute concerning family history of cancer, lung cancer is also included. Many attributes were associated with an increased risk of lung cancer and others a lower risk. As expected, key attributes showing a higher risk included older age, lack of physical activity, COPD, hypertension, other cancers and family history of other cancers, TB and family history of TB and financial status. Attributes associated with lower risk include intense physical activity, younger age, never smokers and higher socioeconomic status. As the results are from univariate linear regression the effect of confounding is apparent. For example, hypertension is associated with age.

The top ten combinations of attributes were selected which showed the best results in identifying lung cancers, out of many thousands of combinations. The selected combinations contained attributes numbering from 7 to 11. The top performing combination included the following attributes: age; activity score; smoking score; any respiratory illness; hypertension; cancer; and Tuberculosis.

We needed to test the performance of the 7-attribute combination henceforth referred to as the Kent & Medway risk prediction tool with the TLHC eligibility criteria. By applying these three criteria to the 30% test population we identified on average 56,663 people (screening cohort) who will be eligible under the TLHC criteria. Among these there were 581 lung cancer cases recorded. We then applied the Kent & Medway risk prediction tool to the same 30% test population, and this predicted a lung cancer risk score for every

individual. From this list we identified the top 56, 663 people and within this population 822 lung cancer cases were recorded. This was on average a benefit of 41.4% over and above the contemporaneous approach.

Table 3: Attributes included in the best performing models and cancer cases detected.

Combination of Attributes	No. of attributes in the Model	Model Runs	Lung Cancer Cases detected	95% CI*	
				Lower	Upper
Age, Activity score, Smoking score, Any respiratory, HT, Cancer, TB	7	100	822	827	817
Age, Active score, Smoking score, Any respiratory, HT, Cancer, TB, Male, Female	9	100	821	826	816
Age, Active score, Smoking score, Gastro, HT, Cancer, Any respiratory, Cancer, Male, Female	10	100	820	825	815
Age, Active score, Smoking score, Gastric condition, HT, Cancer, Any respiratory, TB	8	100	819	824	814
Age, Active score, Smoking score, COPD, Gastric condition, HT, Cancer, Respiratory disease, TB, Male, Female	11	100	818	822	813
Age, Active score, Smoking score, COPD, Gastric condition, HT, Cancer, TB, Male, Female	10	100	817	822	812
Age, Active score, Smoking score, COPD, Endocrine and metabolic condition, Gastric condition, Cancer, TB, Male, Female	10	100	817	822	812
Age, Active score, Smoking score, COPD, Gastric condition, HT, Cancer, Respiratory disease, TB	9	100	817	821	812
Age, Active score, Smoking score, COPD, Endocrine and metabolic condition, Gastric condition, Cancer, TB	8	100	817	821	812
Age, Active score, Smoking score, COPD, Gastric condition, HT, Cancer, TB	8	100	816	821	812

* Confidence interval

Discussion

Our study is an attempt to develop a lung cancer risk prediction tool to identify sections of the population at a higher risk of developing lung cancer. We utilised data on social, demographic, lifestyle and clinical features of the individual and used the power of ML to achieve our objective. We initially identified 16 attributes that could predict the population at a higher risk of lung cancer. Our objective was to increase the power of cancer detection in a defined population as the current targeted TLHC eligibility criteria³² are too broad and blunt. By running simultaneous models using boot strapping we were able to test numerous combinations of attributes running into tens of thousands of model runs which provided us with the best model with 7 attributes. We

adopted a linear regression model which is different to others who have employed a suite of models^{33,34} in lung cancer prediction literature. This is because our objective was to identify a cohort of people at higher risk of lung cancer so that they can be targeted for screening. There is a linear association with many known attributes and risk of lung cancer. Furthermore, lung cancer risk score which is our main outcome of interest is a continuous variable and hence logistic regression is not applicable here. Use of ML has been proposed and adopted in reading computer tomography images³⁴. However, in our study we used data points derived from routine linked administrative data sets which contained information on every patient irrespective of their clinical characteristics to predict their risk of lung cancer by exploiting the potential of ML.

Clinical utility of the work: The product of this work has immediate clinical implications and thus has the potential to improve patient care and resource utilization. As the model outperforms the standard wider TLHC eligibility criteria this would help us to detect up to 40% more cancers. Currently we are exploring how best to incorporate this as a screening and early diagnosis intervention. There are two options under consideration: provide a more comprehensive and refined screening algorithm based on our risk tool compared to that of the TLHC eligibility criteria; and the GP calculates the risk score for each patient during a consultation, similar to Framingham cardiac risk score³⁵ and use this for further action. Using the first option, we can further refine the risk group for screening there by increasing cancer detection and saving scarce cancer diagnostic and treatment resources.

Strengths: We used a place based linked data set entirely produced by a local health system whose primary use was for commissioning intelligence and health care planning purposes. It has the power of painting the entire picture of the population as it contains information from general practice, community health services, mental health services and hospital services. Furthermore, it also included patient level information on key socioeconomic factors and the extent of deprivation, based on integrated well established risk scoring tools. This makes it a powerful repository to develop any risk prediction tool compared to tools that only rely on electronic patient clinical records³⁶. Our data is complete compared to Callender *et al.*³⁷ where there are large number of missing values. We generated relative risks at a very granular level of detail in order to develop our aggregated sixteen attributes. We established a powerful partnership of cancer clinicians, Public Health physicians, epidemiologists, ML experts and leaders from the cancer alliance who were involved throughout from the inception of the project to its completion. This helped us to incorporate varying perspectives. Key stakeholders' views were constantly sought and acted upon during this work. These included regular meetings with the early diagnosis team, digital cancer alliance board, shared health and care analytics board and regional applied research consortium digital innovation group. Patients and the public are represented in most of these in order to ensure that there is support for this initiative.

Limitations:

A few limitations of our study need to be acknowledged. Ethnicity was not included in the model because the data was incomplete. We did our best to locate ethnicity of an individual from various sources still there were large gaps in the data. In future we will ensure that ethnicity is included in further work Data included in the study is only up to 2019. Due to changes in commissioning arrangements, the KID was rendered static and was not updated after this time. We do not anticipate any weakening of the power of the prediction tool due to non-inclusion of more recent data. As we explained earlier our modelling was restricted to linear regression

and did not involve other modelling approaches. This study was undertaken in Kent & Medway in the southeast of England. Hence the question of generalisability across the United Kingdom needs to be considered. In our view it is unlikely that the population and the strength of association between the attributes and lung cancer is so different that the results will not be applicable. However, this may not be true for an international comparison.

Conclusion

In this paper we have demonstrated the exceptional application and utility of ML in developing a risk score for lung cancer using a large, place based linked data set. We involved multidisciplinary stakeholders throughout this work including patients and the public. Our risk prediction tool is superior to the eligibility criteria currently in use in the pilot sites for the TLHC Programme. This is a good example where local experts in fields as diverse as AI, ML, clinical oncologists, and Public Health physicians came together to produce an innovative solution to improve patient care and save scarce health care resources.

Declarations

Data availability: The data is not publicly available as the KID contains pseudonymised person level linked data. However, access to data can be requested via the SHcAB.

Acknowledgements: We are grateful to the SHcAB for granting us permission to access and use the data. We acknowledge the support of Kent & Medway Cancer Alliance. Our sincere thanks to Dr Anjan Ghosh, Director of Public Health, Kent County Council for his support and encouragement.

Funding: The First and second authors received funding from the Kent and Medway cancer alliance to undertake the analysis.

Author Information:

David Howell MPhil, BSC(Hons), MCTS

Director for Quantum Analytica/Data Scientist

Director for Insight and Analytics, Surrey Heartlands Integrated Care System

Ross Buttery

Director for Quantum Analytica/Data Scientist

Padmanabhan Badrinath MD, PhD, MPH

Consultant in Public Health Medicine, Kent County Council

Affiliated Assistant Professor University of Cambridge, UK

Abraham George MBBS, MPH

Consultant Public Health, Kent County Council

Senior Lecturer, Kent and Medway Medical School, UK

Rithvik Hariprasad

Final year engineering student

Vellore Institute of Technology, Vellore, Tamil Nadu, India

Ian Vousden

Interim Managing Director

Thames Valley Cancer Alliance

NHS England - South East

Tina George - BM, MSc

Clinical Director for Cancer, NHS Sussex Integrated Care Board

Clinical Co-Director, Targeted Lung Health Checks Sussex

Early Diagnosis Lead, Kent & Medway Cancer Alliance

Cancer Research UK GP

Cathy Finnis, B.Med.Sci, BMBS, MA

Programme Lead for Early Cancer Diagnosis and Cancer Health Inequalities

Kent and Medway Cancer Alliance

Contributions: All authors contributed to the publication according to the ICMJE guidelines for authorship. All authors read and approved the submitted version of the manuscript. Each author has agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature. Study concept and design: DH, RB, AG. Acquisition of the data: DH, RB, AG. Analysis and interpretation of data: DH, RB, AG, PB, RH, IV, CF, TG. Drafting of the manuscript: DH, RB, PB, AG, RH, IV, CF, TG. Statistical analysis: DH, RB, PB, AG, RH. Manuscript review and approval: DH, RB, PB, AG, RH, IV, CF, TG. Obtained funding: DH, IV.

Ethics declarations: Ethical approval was not required as this work was undertaken as part of the authors' job role and as a service activity to inform health care planning and delivery.

Competing interests: DH and RB are directors of Quantum analytica, a Berkshire based data intelligence company covering the whole of the UK, working with a variety of health and social care providers including the

NHS and local councils. Other authors have no interest to declare.

Additional Information: None

Supplementary information: None

Rights and permissions: We obtained permission from SHCAB to access and use the data.

References

1. Torre LA, Siegel RL, Jemal A. Lung cancer statistics. Lung cancer and personalized medicine: current knowledge and therapies. 2016:1–9.
2. Aggarwal A, Lewison G, Idir S, Peters M, Aldige C, Boerckel W, Boyle P, Trimble EL, Roe P, Sethi T, Fox J. The state of lung cancer research: a global analysis. *Journal of Thoracic Oncology*. 2016 Jul 1;11(7):1040-50.
3. Cancer Research UK. Lung Cancer Statistics. Cancer Research UK. [Internet]. Accessed 2023 June 8. Available from: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer>
4. Wynder EL, Hoffmann D. Smoking and lung cancer: scientific challenges and opportunities. *Cancer Research*. 1994 Oct 15;54(20):5284–94.
5. Chaitanya Thandra K, Barsouk A, Saginala K, Sukumar Aluru J, Barsouk A. Epidemiology of lung cancer. *Contemporary Oncology/Współczesna Onkologia*. 2021;25(1):45–52. doi:10.5114/wo.2021.103829.
6. National Institute of Health and Care Excellence. Suspected cancer: recognition and referral. NICE guideline [NG12] Published: 23 June 2015 Last updated: 15 December 2021. Internet. <https://www.nice.org.uk/guidance/ng12>. Accessed June 2023.
7. Samson P, Patel A, Garrett T, Crabtree T, Kreisel D, Krupnick AS, Patterson GA, Broderick S, Meyers BF, Puri V. Effects of Delayed Surgical Resection on Short-Term and Long-Term Outcomes in Clinical Stage I Non-Small Cell Lung Cancer. *Ann Thorac Surg*. 2015 Jun;99(6):1906–12.
8. McPhail S, Johnson S, Greenberg D, Peake M, Rous B. Stage at diagnosis and early mortality from cancer in England. *British journal of cancer*. 2015 Mar;112(1):S108-15.
9. Cancer Research UK. Why is early diagnosis important? [Internet]. Available from: <https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important>. Accessed 18th June 2023.
10. NHS. Chapter 3 Further progress on care quality and outcome. NHS Long Term Plan. Internet. <https://www.longtermplan.nhs.uk/online-version/chapter-3-further-progress-on-care-quality-and-outcomes/better-care-for-major-health-conditions/cancer/> Accessed on 22 June 2023.
11. Corral J, Espinàs JA, Cots F, Pareja L, Solà J, Font R, Borràs JM. Estimation of lung cancer diagnosis and treatment costs based on a patient-level analysis in Catalonia (Spain). *BMC health services research*. 2015 Dec;15:1–0.
12. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism*. 2017 Apr 1;69:S36-40.
13. Chiu HY, Chao HS, Chen YM. Application of artificial intelligence in lung cancer. *Cancers*. 2022 Mar 8;14(6):1370.

14. Hindman M. Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*. 2015 May;659(1):48–62.
15. Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK. Lung cancer risk prediction: a tool for early detection. *International journal of cancer*. 2007 Jan 1;120(1):1–6.
16. Public Health England. NHS population screening: care pathways [Internet]. August 2021 [cited 2023 May 26]. Available from: <https://www.gov.uk/government/collections/nhs-population-screening-care-pathways>
17. Bobrowska A, Murton M, Seedat F, Visintin C, Mackie A, Steele R, Marshall J. Targeted screening in the UK: A narrow concept with broad application. *The Lancet Regional Health-Europe*. 2022 May 1;16:100353.
18. Crosbie PA, Balata H, Evison M, Attack M, Bayliss-Brideaux V, Colligan D, Duerden R, Eaglesfield J, Edwards T, Elton P, Foster J. Second round results from the Manchester 'Lung Health Check' community-based targeted lung cancer screening pilot. *Thorax*. 2019 Jul 1;74(7):700-4.
19. Crosbie PA, Balata H, Evison M, Attack M, Bayliss-Brideaux V, Colligan D, Duerden R, Eaglesfield J, Edwards T, Elton P, Foster J. Implementing lung cancer screening: baseline results from a community-based 'Lung Health Check' pilot in deprived areas of Manchester. *Thorax*. 2019 Apr 1;74(4):405-9.
20. Office for National Statistics. Population and Migration - Population Projections. ONS. [Internet]. Accessed 2023 June 18. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections>
21. Macmillan Cancer Support. 2022 Cancer Statistics Factsheet. [Macmillan.org.uk](https://www.macmillan.org.uk). [Internet]. Accessed 2023 June 18. Available from: <https://www.macmillan.org.uk/dfsmedia/1a6f23537f7f4519bb0cf14c45b2a629/9468-10061/2022-cancer-statistics-factsheet>
22. Lewer D, Bourne T, George A, Abi-Aad G, Taylor C, George J. Data resource: the Kent integrated dataset (KID). *International Journal of Population Data Science*. 2018;3(1).
23. Statistical Bulletin. 2021 Mid-Year Population Estimates: Age and Sex Profile. Kent Analytics. January 2023. Available online: https://www.kent.gov.uk/__data/assets/pdf_file/0019/14725/Mid-year-population-estimates-age-and-gender.pdf (accessed on 23 March 2023)
24. Health & Social Care Maps. PDF Social Care Maps. KPHO. [Internet]. Accessed 2023 June 1. Available from: <https://www.kpho.org.uk/joint-strategic-needs-assessment/health-and-social-care-maps/pdf-social-care-maps>
25. Kent and Medway Public Health Observatory. Annual Public Health Report - APhR 2021. KPHO. [Internet]. Accessed 2023 June 1. Available from: https://www.kpho.org.uk/__data/assets/pdf_file/0003/138270/Kent-APHR-2021-Coastal-Communities.pdf
26. Potdar K, Pardawala TS, Pai CD. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*. 2017 Oct;175(4):7–9.
27. Carr LL, Jacobson S, Lynch DA, Foreman MG, Flenaugh EL, Hersh CP, Sciruba FC, Wilson DO, Sieren JC, Mulhall P, Kim V. Features of COPD as predictors of lung cancer. *Chest*. 2018 Jun 1;153(6):1326-35.

28. Tenkanen L, Teppo L, Hakulinen T. Smoking and cardiac symptoms as predictors of lung cancer. *Journal of Chronic Diseases*. 1987 Jan 1;40(12):1121-8.
29. Quang Hung Nguyen, Hai-Bang Ly, Lanh Si Ho, Nadhir Al-Ansari, Hiep Van Le, Van Quan Tran, Indra Prakash, Binh Thai Pham, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil", *Mathematical Problems in Engineering*, vol. 2021, Article ID 4832864, 15 pages, 2021. <https://doi.org/10.1155/2021/4832864>
30. Python [Internet]. Available from: <https://www.python.org/about/> Accessed 22 June 2023
31. Marcus MW, Field JK. Is bootstrapping sufficient for validating a risk model for selection of participants for a lung cancer screening program?. *Journal of Clinical Oncology*. 2017 Mar 10;35(8):818–9.
32. Lung health checks in Kent. Internet. <https://www.kentandmedway.icb.nhs.uk/your-health/local-services/kent-and-medway-cancer-alliance/lung-checks> accessed on 22 June 2023.
33. Dritsas E, Trigka M. Lung Cancer Risk Prediction with Machine Learning Models. *Big Data and Cognitive Computing*. 2022 Nov 15;6(4):139.
34. Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. *Translational lung cancer research*. 2018 Jun;7(3):304.
35. MDCalc. Framingham Risk Score (Hard Coronary Heart Disease). [Internet]. Available from: <https://www.mdcalc.com/calc/38/framingham-risk-score-hard-coronary-heart-disease>. [Accessed: June 20, 2023].
36. Raghu VK, Walia AS, Zinzuwadia AN, Goiffon RJ, Shepard JA, Aerts HJ, Lennes IT, Lu MT. Validation of a Deep Learning–Based Model to Predict Lung Cancer Risk Using Chest Radiographs and Electronic Medical Record Data. *JAMA Network Open*. 2022 Dec 1;5(12):e2248793.
37. Callender T, Imrie F, Cebere B, Pashayan N, Navani N, van der Schaar M, Janes SM. Assessing eligibility for lung cancer screening: Parsimonious multi-country ensemble machine learning models for lung cancer prediction. *medRxiv*. 2023 Jan 29:2023–01.

Table

Table 2 is available in the Supplementary Files section.

Figures

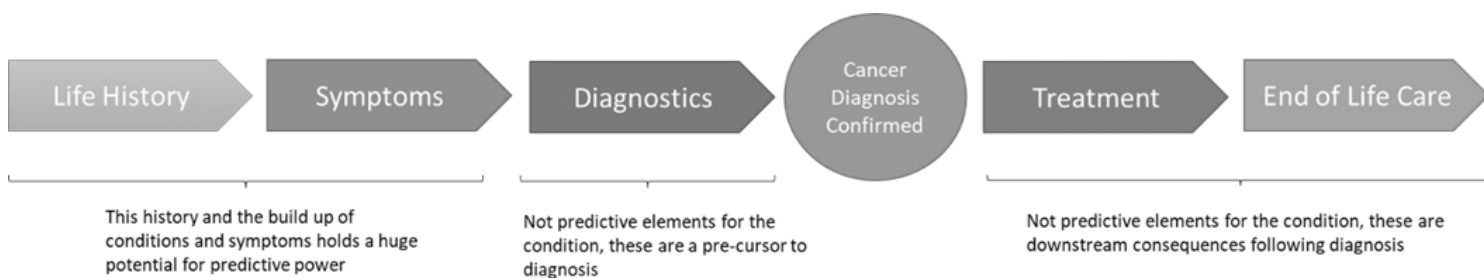


Figure 1

Pathways leading up to and beyond a Lung Cancer Diagnosis for patients.

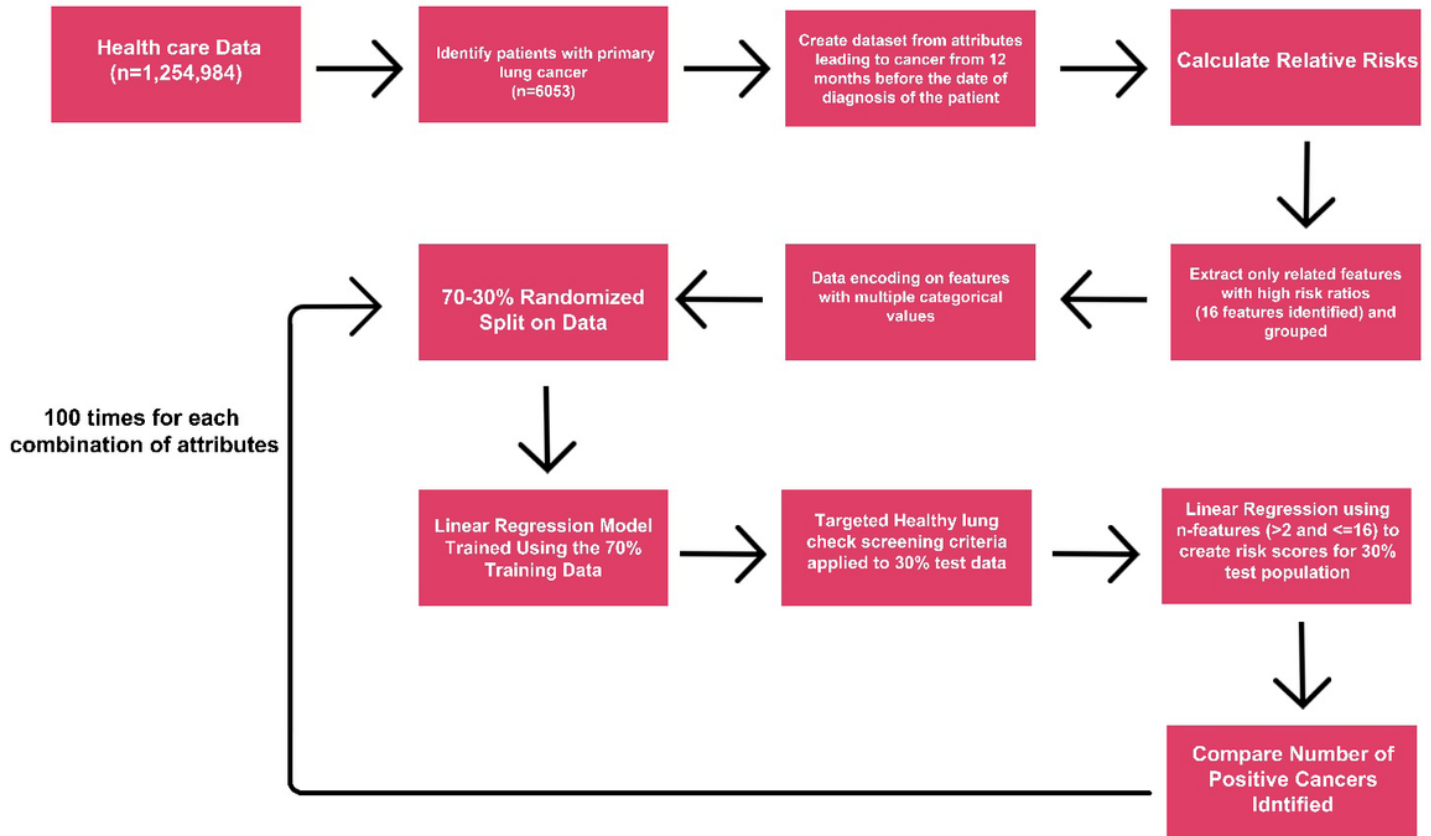


Figure 2

shows the steps involved from the beginning to the end of the study process. This spans from extracting relevant data from the KID to comparing the number of lung cancer cases detected using the most successful model and the criteria used in the TLHC programme.

Supplementary Files

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

- [Table2.docx](#)