Plasma Protein Biomarkers for Early Prediction of Lung Cancer

Michael P.A. Davies, PhD,[a] Takahiro Sato, PhD,[b] Haitham Ashoor, PhD,[b] Liping Hou, PdD,[c] Triantafillos Liloglou, PhD,[d] Robert Yang, PhD,[b] John K. Field PhD, FRCPath [a]*

[a] *Department of Molecular and Clinical Cancer Medicine, Institute of Systems, Molecular & Integrative Biology, The University of Liverpool,* William Henry Duncan Building, 6 West Derby Street, *Liverpool* L7 8TX*, UK*

[b] *World Without Disease Accelerator, Johnson & Johnson, 10th Floor 255 Main St, Cambridge, MA 02142, USA*

[c] *Population Analytics & Insights, Data Science, Janssen R&D, 1400 McKean Rd, Spring House, PA 19477, USA*

[d] *Faculty of Health, Social Care & Medicine, Edge Hill University, St Helens Road, Ormskirk, Lancashire L39 4QP, UK*

Joint first authors: MD, TS & HA

*** Corresponding author**

Address for correspondence:

Professor John K Field,
Dept. of Molecular and Clinical Cancer Medicine,
Institute of Systems, Molecular & Integrative Biology,
University of Liverpool,
William Henry Duncan Building,
6 West Derby Street,
L7 8TX,
Liverpool, UK
Email: J.K.Field@liverpool.ac.uk

SUMMARY

Background

Individual plasma proteins have been identified as minimally invasive biomarkers for lung cancer diagnosis with potential utility in early detection. Plasma proteomes provide insight into contributing biological factors; we investigated their potential for future lung cancer prediction.

Methods

The Olink® Explore-3072 platform quantitated 2941 proteins in 496 Liverpool Lung Project plasma samples, including 131 cases taken 1-10 years prior to diagnosis, 237 controls, and 90 subjects at multiple times. 1112 proteins significantly associated with haemolysis were excluded. Feature selection with bootstrapping identified differentially expressed proteins, subsequently modelled for lung cancer prediction and validated in UK Biobank data.

Findings

For samples 1-3 years pre-diagnosis, 240 proteins were significantly different in cases; for 1-5 year samples, 117 of these and 150 further proteins were identified, mapping to significantly different pathways. Four machine learning algorithms gave median AUCs of 0·76-0·90 and 0·73-0·83 for the 1-3 year and 1-5 year proteins respectively. External validation gave AUCs of 0·75 (1-3 year) and 0·69 (1-5 year), with AUC 0·7 up to 12 years prior to diagnosis. The models were independent of age, smoking duration, cancer histology and the presence of COPD.

Interpretation

The plasma proteome provides biomarkers which may be used to identify those at greatest risk of lung cancer. The proteins and the pathways are different when lung cancer is more imminent, indicating that both biomarkers of inherent risk and biomarkers associated with presence of early lung cancer may be identified.

Funding

Janssen Pharmaceuticals Research Collaboration Award; Roy Castle Lung Cancer Senior Research Fellowship.

Keywords: lung cancer prediction; early-detection, plasma, proteins, proteomics

Research in context

Evidence before this study

Differences in specific plasma protein levels have been previously shown to be indicative for lung cancer diagnosis, or related to imminent lung cancer. However, more comprehensive plasma protein profiling over longer time periods pre-diagnosis has not been studied.

Added value of this study

The findings in this paper have confirmed the predictive power of plasma protein profiling for prediction of future lung cancer diagnosis, identifying potential protein biomarkers for early detection. That biomarker proteins selected using longer pre-diagnostic time points only partially overlap those selected using samples from later time points, and represent different molecular pathways, suggests that both biomarkers for inherent cancer risk and occult tumour detection can be identified. This is further supported by the differing longitudinal levels across multiple time points, including at diagnosis.

Implications of all the available evidence

When considering the value and utility of plasma protein biomarkers, future studies should consider the relationship between plasma levels and the possibility of undiagnosed tumours or the association of protein levels with biological manifestation of risk of future cancer. Whilst the former facilitates earlier diagnosis, the latter contributes to the stratification of high-risk individuals for screening, or targeted preventative measures.

Introduction

Lung cancer continues to be the largest cause of cancer deaths worldwide, representing 18% of all cancer deaths, an estimated 2·2 million new cancer cases and 1·8 million deaths.[1] This is closely related to the fact that most lung cancers are detected at a late stage,[2] when more effective treatment options such as surgery are not available and outcomes are poor.[3] Low-dose CT screening has been shown to detect a higher proportion of early-stage disease than traditional symptomatic diagnosis, leading to improved outcomes.[4-6] However, it is currently limited to those with a significant history of smoking, since this is used as a key decision-making factor for screening. However, selection for LDCT on the basis of smoking and age alone, does not fully account for the differing risk amongst smokers due to genetic or environmental factors.[7]

Lung cancer risk assessment (based not just on smoking, but other demographic variables and medical history) attempts to identify those most likely to suffer from lung cancer in the future, e.g. the next 5-6 years. Effective tools are available and have proved useful, e.g. in identification of those who benefit most from lung cancer screening, contributing to cost effectiveness of diagnostic intervention.[5,7] However, many lung cancers are still missed, either because they are in those currently excluded from screening, or because screening uptake is sub-optimal. Biological assessment of risk, utilising minimally invasive biomarkers, may widen the applicability of LDCT screening, or encourage greater uptake by re-enforcing personal risk awareness.

Lung cancer risk biomarkers (indicating a predisposition to lung cancer) overlap significantly with diagnostic biomarkers (indicating a likelihood of current disease). A variety of different biomarkers have been demonstrated to aid early diagnosis of lung cancer, either alongside [8] or in the absence of LDCT screening,[9] but there are still unmet clinical needs [10] and technical challenges. Nevertheless, the addition of diagnostic plasma protein biomarkers has been demonstrated to improve current risk scores [11,12] and new, more comprehensive discovery platforms provide further opportunities to identify further biomarkers for lung cancer risk prediction. Risk assessment may also help to identify those who will benefit from preventative medical intervention; plasma proteins may be particularly advantageous, as they provide biological insights into the potential preventative treatment approaches.[13]

Here we have identified a case-control cohort from the Liverpool Lung Project observational study that includes subjects with samples taken 1 or more years prior to their lung cancer diagnosis; with controls matched for age, sex and smoking history. We identify plasma proteins significantly associated with lung cancer status, build models predicting lung cancer diagnosis and validate them in an independent observational dataset.

Methods

LLP cohort

Samples, risk questionnaires and clinical data from Liverpool Lung Project (LLP) participants [14] were obtained following voluntary informed consent, in accordance with the Declaration of Helsinki. Ethical approval was obtained from the Liverpool Central Research Ethics Committee (ref 97/141). Lung cancer cases were identified through NHS Digital (now NHS England) via the National Cancer Registration and Analysis Service or through case note review. Individual level healthcare data is available only for restricted purposes and must be governed by a data sharing agreement, so is only available upon request.

EDTA plasma samples from LLP subjects were collected by standardised protocols (between 1998 and 2016), with a single cell depletion centrifugation (2200g, 15 min.) prior to storing at -80^{o}C and a further cell depletion spin after thawing, before being aliquoted for Olink studies and refrozen for shipment.

The cases and controls for this study were selected retrospectively as a nested case-control cohort from the LLP population cohort (Supplemental Figure S1). LLP population cohort subjects without lung cancer at the time of recruitment, but with subsequent diagnosis of primary lung cancer within 5 years were identified for the primary discovery cohort (cases, Table 1). Non-small cell lung cancer cases included almost equal numbers of adenocarcinoma (n= 53) and squamous cell carcinoma (n= 49) and were either early stage (45%) or late stage (52%) at the time of diagnosis (Supplementary Table S1). Samples at diagnosis (n=23), 1-3 years prior to diagnosis (n=21), 3-5 years prior to diagnosis (n=30) or 5-10 years prior to diagnosis (n=33), were identified for longitudinal studies from 42 cases (Table S2), along with 110 longitudinal samples at the same time points from 48 controls.

For each case, sex (self-reported as sex assigned at birth) and age at plasma sample were used to match control subjects (2 per case for discovery cohort and 1 per case for longitudinal studies). Controls were also selected to have the same smoking status (current/former/never) at the time of sampling and similar lifetime smoking duration (based on all forms of tobacco). Where multiple longitudinal bio-specimens were available from cases, controls were identified with multiple samples at approximately the same intervals. Most subjects were smokers at the time of initial blood collection, with only 10 never smokers, but 24 had quit smoking at the time of the last sample used.

Olink platforms

Pre-diagnosis plasma proteomics was assessed in a cross-sectional sub-cohort (292 subjects, 1-5 years before diagnosis), and a longitudinal sub-cohort (246 samples from 144 subjects, 5-10 years before diagnosis, 2-5 years before diagnosis, and at time of diagnosis). We generated plasma proteomics data using the Olink Explore 3072 platform (2941 proteins), which consists of 8 separate panels: Oncology, Oncology II, Cardiometabolic, Cardiometabolic II, Inflammation, Inflammation II, Neurology, and Neurology II. We generated PCA plots with all proteins and samples, and filtered 6 samples with > 5 standard deviations from the mean. We then generated PCA for each panel separately, and filtered an additional 5 samples with > 5 standard deviations from the mean. Data was also generated using the Olink Target 96 platform (panels: Cardiometabolic, Cardiovascular II, Cardiovascular III, Cell Regulation, Development, Immune Response, Inflammation, Metabolism, Neuro Exploratory, Neurology, Oncology II, Oncology III, Organ Damage).

Haemolysis

Haemolysis is known to contribute to increased levels of some proteins in plasma [15]. To avoid potential false-positives results due to haemolysis-associated signals, we systematically removed proteins that were found to be significantly associated with haemolysis (Supplementary Table S3). Each sample in the LLP cohort had a haemolysis score assigned ranging from 0 to 4. A linear model was generated to identify proteins significantly associated with haemolysis, with 1112 proteins out of 2941 proteins measured by Olink Explore identified based on FDR < 0·01. These proteins were filtered out from further analysis.

UK Biobank

Olink data was generated in UK Biobank (UKB) data as part of the UK Biobank Pharma Proteomics Project [16]; the UK Biobank population (age 40 to 69 years) is younger than the LLP population (age 48 to 84 years) . We analysed the initial batch of data which was generated using the Olink Explore 1536 platform (1472 proteins) on 54,306 UKB participants. We extracted future cancer cases from UK Biobank cancer registry. We defined lung cancer cases using the ICD10 code of C34. Cancer cases are restricted to the first occurrence, have future cancer from the baseline blood draw, and have Olink data. After applying our selection criteria, the total number of cases is 392 (Supplementary Figure S2, Supplementary Table S4).

Controls were defined as individuals with no record of cancer, who did not self-report any previous cancer incidents, and if deceased cancer was not the cause of death. We matched controls to cancer cases by age, sex, smoking status and race, using the K-nearest neighbour method to generate matching controls. Two patient-to-control ratios were implemented; one is a balanced ratio where the ratio of cancer to control is 1:1, and the second represents the risk of getting lung cancer as 1 cancer : 14 controls (392 cases and 5500 controls).

For pan-cancer analysis, we repeated the above process for each cancer type then we combined control samples from different cancer types into one pooled control sample; ICD 10 cancer codes: Prostate, C61; Breast, C50; Colorectal, C18 & C19; Uterine Cancer, C44; Kidney Cancer, C64; Pancreatic, C25; Bladder, C67; Stomach, C16; Liver, C22.

Machine Learning

Feature selection was performed on the discovery cohorts (Table 1) by bootstrapping differential expression on a random set of 50% of the dataset 1000 times using a linear model with age, sex, and pack years as covariates, and significant proteins were defined as being differentially expressed between cases and controls (P < 0·05 linear model anova) at least 100 times. Proteins associated with haemolysis were then filtered out. Four different machine learning algorithms were trained [Elastic Net,[17] Random Forest,[18] Support Vector Machine,[19] XGBoost [20]] as a binary model to predict cancer vs. control either at 1-3 years before diagnosis or 1-5 years before diagnosis of lung cancer. Receiver operating characteristic area under the curve values (AUCs) from the models are reported as the median AUC from 5-fold cross validation repeated 5 times. To predict future cancer in UKB individuals, we intersected selected proteins with proteins available in UKB data and trained Support Vector Machine (SVM) classifiers using this set of proteins.

Pathway Enrichment

For GO biological process pathways gene set enrichment, 7658 gene sets were downloaded from msigdb (www.gsea-msigdb.org), and the list was filtered to only include proteins measured by the Olink Explore platform (2941 proteins). Hypergeometric tests were performed separately on proteins higher or lower in lung cancer cases from the 1-3Y and 1-5Y models, with the background as the 2941 proteins measured by Olink.

Role of the funding source

This work was supported by Janssen Pharmaceuticals through a Research Collaboration Agreement with the University of Liverpool. Writing of the manuscript and the decision to submit it for publication was a joint undertaking by members of the research team from both parties. Scientists from both parties were involved in study design, data collection from analytical tests, analysis of anonymous data, and interpretation of results, but only University of Liverpool researchers were responsible for patient recruitment or identifiable personal healthcare data.

The Liverpool Lung Project has also been funded by the Roy Castle Lung Cancer Foundation, but the they had no direct role in production of this manuscript.

Results

Lung cancer associated plasma proteins

We combined patient samples taken 1-3 years before diagnosis (1-3Y) from the cross-sectional and longitudinal sub-cohorts to build models to predict development of future lung cancer. We identified 422 proteins that were differentially expressed between healthy subjects and future lung cancer cases 1-3Y prior to diagnosis. 240/422 proteins were kept for further analysis (158 up in cases and 82 down) after filtering out proteins that were significantly associated with haemolysis (Supplementary Table S3). A subset of these proteins was also measured on the Olink Target 96 platform and these correlated well with the Olink Explore platform; 262/265 of the overlapping proteins had a significant correlation with FDR < 0·05 (Supplemental Figure S3 and Supplemental Table S6).

Machine Learning

Training four different machine learning algorithms on the LLP cohort (Elastic Net, Random Forest, Support Vector Machine, XGBoost, 5-fold cross validation repeated 5 times) using the 240 proteins in the 1-3Y cohort generated median AUCs from the cross validation ranging from 0·76 to 0·90 (Figure 1a).

Combined z scores were generated from the differentially expressed proteins at 1-3Y before diagnosis and were plotted over time, including additional longitudinal samples (Figure 1b). The difference between cases and controls was greater closer to diagnosis. The 1-3Y combined z score differentiated between controls and cases at 1-3 years before diagnosis, but not at 3-5 years or 5-10 years before diagnosis. Individual patient trajectories of the combined z scores indicate that patients that developed cancer were more likely to have an upward trajectory of their z score over time (Supplemental Figure S4).

The combined z scores did not differ between stage of cancer at time of diagnosis (Figures 2a). The only significant difference between stages was at 5-10 years before diagnosis, where it was higher for stage I than stage IV. However, at this time point the healthy and lung cancer z-scores didn't demonstrate a difference overall. The combined z scores also did not correlate with pack years regardless of time before diagnosis, whether looking at healthy or lung cancer subjects (Figures 2b). The z score had a stronger signal in squamous cell carcinoma 3-5 years before diagnosis, had no correlation with age in pre-diagnostic samples, and had no association with diagnosis of COPD (Supplemental Figure S5).

UK Biobank validation

These 1-3Y trained model was tested on samples in the UK Biobank using SVM, which was the model that had the best performance in the training cohort. Only proteins that were measured in both LLP and UKB were used in the models since the UKB cohort measured a smaller panel of proteins using the Olink Explore platform: 107/240 for the 1-3Y model. We constructed a UK biobank cohort that includes 392 future lung cancer cases and 5500 cancer-free controls (see Methods). Our 1-3Y model proteins gave an AUC from the cross validation of 0·75 for predicting cancer 1-3Y before diagnosis (Figures 1c). An AUC of ~0·7 was retained for predicting cohorts that included patients 12 years prior to diagnosis (Figure 1d). We also demonstrate that our model is highly specific to lung cancer; when we try to predict other types of cancer using our model AUCs were around 0·5 (Figure 1e). Sub-cohort analysis indicated that the model retained performance in non-smokers, patients younger than the age from the recommended screening guidelines and both sexes (Table 2); it also performed similarly for different histological subtypes (Supplementary Table S7).

Longer term prediction

We further investigated the ability of plasma proteins to predict lung cancer by repeating our analysis using sample taken 1-5 years (1-5Y) prior to diagnosis and matched controls. We identified 489 proteins 1-5Y before diagnosis that were differentially expressed between future lung cancer and healthy subjects. After filtering out proteins that were significantly associated with haemolysis, 267/493 proteins were kept for further analysis (119 up in cases and 148 down). Of these, 117 of these were also identified for the 1-3Y analysis, 69 up in cases and 48 down in cases (Supplementary Table S5). Hence, over half of those plasma proteins significantly altered in the future lung cancer cases 1-5Y before diagnosis were not identified as significantly altered 1-3Y before to diagnosis (n = 150, 50 up in cases and 100 down in cases).

The combined z score for the 1-5Y proteins had the same relationship to histology, COPD (Figure S5) and smoking pack year histology as the 1-3Y proteins. However, in contrast to 1-3Y proteins (Figure

1b), the 1-5Y combined z score differentiated between controls and cases at both 1-3Y and 3-5Y before diagnosis (Figure 3b), had no relationship to stage (Figure S5f) and had a negative correlation with age in pre-diagnostic cancer cases and healthy controls (Figure 3c).

Training four different machine learning algorithms (with 5-fold cross validation repeated 5 times) using these 267 1-5Y proteins generated median AUCs from the cross validation ranging from 0·73 to 0·83, as shown in Figure 3a. During external validation, the model based on the 129 1-5Y proteins measured in the UKB data gave an AUC of 0·69 for predicting lung cancer 1-5Y before diagnosis, which was not significantly different to the 1-3Y model. As with the 1-3Y model, AUC remained around 0·7 even for samples up to 12 years prior to diagnosis.

Biological pathways

Gene enrichment analysis was performed to investigate potential biological pathways implicated in the risk of future lung cancer, being either increased in plasma (over-represented in cases) or decreased in plasma (under-represented in cases). For the top 20 pathways enriched for proteins either higher or lower in cases, there was limited overlap between 1-3Y and 1-5Y cohorts (Figure 4); only 3 pathways over-represented in cases and 3 pathways under-represented in cases were shared between the 1-3Y and 1-5Y proteins. Of those pathways with higher plasma protein levels in cases, of the 152 pathways with P<0·05 for either cohort, 57 were significant for 1-5Y only, 83 for 1-3Y only and only 12 for both (Supplementary Table S8a). For proteins with lower levels in cases, of the 138 pathways with P<0·05 for either cohort, 55 were significant for 1-5Y only, 74 for 1-3Y only and only 9 for both (Supplementary Table S8b).

That individual proteins may be associated with different aspects of lung cancer risk and/or presence of undetected lung cancer is exemplified by looking at how levels change over time (Supplemental Figure S6) in those cases and controls with longitudinal samples (Table S2). Some increase (e.g. PDIA4, RBPMS2) or decrease (e.g. ENPP6) the closer the sample is taken to diagnosis; others are consistently higher (e.g. CEACAM5) or lower (e.g. MFGE8) varying less over time, but many exhibit a combination of both traits.

Discussion

We have performed comprehensive plasma protein discovery, using the Olink® Explore 3072 platform, on plasma samples from the Liverpool Lung Project (LLP) [14] taken at various times prior to lung cancer diagnosis. This provides insight into early predictive biomarkers and how they change over time. The plasma proteome provides protein biomarkers which may be used to identify those at greatest risk of lung cancer, 5 or more years prior to diagnosis. This approach may provide an opportunity to identify patients who would benefit from novel preventative approaches (for pharmaceutical or vaccination interventions) or who would be eligible for lung cancer screening despite not conforming to current smoking-related selection criteria.

Selecting proteins by bootstrapping differential expression, we identified 425 and 493 proteins respectively in the 1-3Y and 1-5Y cohorts. However, we also noted that many of these proteins were associated with haemolysis. As haemolysis-associated proteins would give potential false positive signals if any healthy samples were haemolysed, and it is possible that haemolysis is more often seen in lung cancer patients than healthy individuals, we chose to remove any proteins that were associated with haemolysis, leaving 240 (1-3Y) and 267 (1-5Y) proteins with each panel combined in a z score to investigate relationships with clinical and epidemiological factors. We found no association with smoking (pack years or duration) or with a history of COPD; a negative association with age was seen for pre-diagnostic samples and controls for the 1-5Y z score only. Hence, the plasma proteins are not directly related to known risk factors for cancer, meaning they are more likely to provide additional useful information when used in conjunction with lung cancer risk scores and be unrelated to smoking-induced inflammation. Furthermore, there was no association with stage of disease at diagnosis (apart from the 1-3Y z score association with early stage, albeit at 5-10 years pre-diagnosis, when not significantly different to control samples) and only a weak association with histological type specifically at 3 - 5 years before diagnosis. These results indicate that the identified proteomic signals are likely to be useful for prediction of any sub-type of non-small cell lung cancer, regardless of stage.

Although we have identified 240 plasma proteins differentially expressed 1-3 years prior to diagnosis and 267 proteins 1-5 years prior to diagnosis, only 117 of the total 390 proteins (30%) were identified in both analyses. This is noteworthy, as the plasma proteome reflects not just the presence of an occult, pre-diagnosis tumour (with signals most likely closer to diagnosis), but immune response to pre-malignant disease and the biological response to inflammation associated smoking and environmental factors (risk factors that are not necessarily higher at time of diagnosis). Furthermore, when mapped on to pathways by gene set enrichment analysis, there was limited overlap between the top pathways from 1-3Y and 1-5Y (only 21 pathways of 290 with significant enrichment), indicating different biological pathways drive the signal for long-term and short-term risk. Pathway analysis provides valuable insight into potential biological mechanisms underpinning the differential expression, potentially providing insights into targets for preventative treatment for those at high risk of lung cancer. However, it should be noted that although extensive, the Olink panels have been carefully curated to reflect specific pathways. Hence these will be over-represented in any subsequent pathway analysis, potentially biasing results.

As might be expected, the z score based on those selected based on 1-5Y samples showed a greater differential expression at 3-5 years prior to diagnosis than that based on 1-3Y protein. Nevertheless, four different machine learning algorithms demonstrated that both the 1-3Y and 1-5Y proteins were able to predict lung cancer up to 5 years prior to diagnosis (AUCs of 0·76-0·90 for the 1-3Y models and 0·73-0·83 for the 1-5Y models). Remarkably, in the UK Biobank validation it was shown that either set of proteins were able to predict lung cancer to the same extent (AUC = 0·7) up to 12 years prior to diagnosis. It is important to note that this cancer prediction was exclusive to lung cancers, with other future cancers in the UK Biobank cohort not predicted, indicating that both the predisposing factors and the tumour-released proteome are likely distinctive for different tumours. Furthermore, in the UK Biobank validation, we observed that the predictive power was maintained to some extent in never smokers (AUC = 0·62) compared to smokers (AUC = 0·69) and was also predictive in those aged 40-55 years (AUC 0·78), who would not usually be eligible for LDCT lung cancer screening; there was also some evidence that it performed better in males (AUC 0·72) than females (AUC 0·66). It is therefore possible that plasma proteome biomarkers might help to expand lung cancer prediction risk scores for better utility within groups currently excluded from the benefit of LDCT screening. However, this would need to be tested in larger populations of younger subjects and never smokers, as these groups are under-represented in most lung cancer cohorts.

Looking at longitudinal samples, the combined z score for the 1-3Y proteins rises significantly towards diagnosis. However, for the 1-5Y protein, differences extend to earlier in disease progression and the levels of some proteins were not increased to as great an extent closer to diagnosis. This indicates that they may represent marker of risk, being indicative of either genetic predisposition or smoking-related damage, rather than being tumour-released or tumour-reactive proteins. Risk biomarkers, rather than being used for early diagnosis, may allow one to identify those who would benefit most from preventative measures, including therapeutic-prevention. For example, inflammation has been shown to be a potential target when post-hoc analysis of the CANTOS trial of Canakinumab (an anti-interleukin-1β monoclonal antibody), for prevention of recurrent vascular events in patients with a persistent pro-inflammatory response, demonstrated a protective effect on lung cancer incidence and mortality [21]; although subsequent trails in treatment of existing cancers have so far proved inconclusive.

Plasma proteins have been shown to provide a means to predict those most at risk of future lung cancer.[11,12] Similarly, the models we have identified could be considered as candidates for inclusion in risk profiling for LDCT screening, or for expedited referral of symptomatic patients. However, the cost-effectiveness of this approach will have to be validated in the context of LDCT screening and/or primary care referrals. One limitation of our analysis is the retrospective nature of the analysis, although this is unavoidable given the need for sufficient follow-up to identify both the cancer cases and to establish that the controls are cancer free. Retrospective validation of risk prediction in similar prospectively collected cohorts such as UK Biobank, should be supplemented with prospective validation. A further limitation of our analysis is that we did not attempt to reduce our relative protein level analysis to a plasma protein biomarker that might more reasonably be measured cheaply using technologies employed in routine clinical diagnosis. This is likely to be more feasible for a smaller number of proteins and to require absolute (fully quantitative) measurement with external validation at population level. It would also be potentially worthwhile incorporating clinical, lifestyle and environmental risk factors into a joint model. Whilst that requires larger datasets, it is notable that there were no strong positive associations between the plasma protein models we identified and other known risk factors such as age, smoking and COPD.

A number of other biomarker platforms have been used to improve lung cancer diagnosis in asymptomatic patients, with a view towards early detection based on circulating tumour DNA including DNA methylation[22], fragmentomics[23] and genomics approaches.[24,25] Unlike the plasma protein biomarkers described here, these rely predominantly on detecting nucleic acids released from cancer cells into the circulation, the concentration of which increase as tumour grow or spread. This underlies the relatively poor performance seen for some tests in early stage disease (e.g. a sensitivity for Grail test of only 21.9% for stage I, but 79.5% for stage II, 90.7% for stage III and 95.2% for IV).[22] However, it should be noted that Grail is a multi-cancer test and these sensitivities might be improved by careful selection of lung cancer specific methylation biomarkers.[26] Those tests that rely on biomarkers released by cancer cells will always face the challenge of low levels in early stage disease and they are also less likely to reflect inherent cancer risk in the same way as biomarkers that capture host factors (such as smoking-induced damage or immune-response to precursor lesions and early tumours).

We have demonstrated that some proteins are associated with longer-term risks, rather than increasing closer to diagnosis (and presumably either being tumour-released or indirectly associated with tumour burden). Other research groups have also demonstrated associations between specific proteins and lung cancer risk within 1 year of diagnosis,[27] but our discovery study provides data on a much wider range of proteins. Other biomarkers may include contributions from host cells, including lipidomics[28] and miRNA.[29] The lipidomics approach taken by Wang et al. is predicated on altered lipid metabolism in cancer cells and identified levels of 9 lipids in plasma that were highly accurate for lung cancer, predominantly early stage, in an independent LDCT screening cohort (more than 90% sensitivity and 92% specificity). Profiling miRNA has been validated to an even greater extent in Italian LDCT screening studies, including for diagnosis[30] and personalising screening intervals.[31]

By contrast our results are preliminary and we did not attempt to reduce our relative protein level analysis to a plasma protein biomarker that could be tested cost effectively using technologies employed in routine clinical diagnosis. This is potentially feasible for a smaller number of validated proteins and requires absolute (fully quantitative) measurement with external validation at population level (as described by the INTEGRAL consortium).[32] Nevertheless, our data adds significantly to such studies, providing longer-term data and longitudinal data for individuals. One potential advantage highlighted is that the plasma proteome could provide insight both into the individual's risk of lung cancer and a means of detecting imminent early stage disease.

In conclusion, the plasma proteome analysis, performed on pre-diagnostic samples from lung cancer patients and lung cancer free controls, identified two partially overlapping panels of proteins from samples 1-3 years or 1-5 years prior to cancer. These panels mapped to predominantly different pathways, but both were predictive for lung cancer on internal and external validation. That samples further from diagnosis displayed different patterns of predictive plasma proteins may indicate that they reflect biological risk, rather than tumour-associated changes. The latter are nevertheless significant in both panels, the combined z scores of which are highest at diagnosis.

Contributors

Michael Davies: Conceptualization, Methodology, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration, Funding acquisition

Takahiro Sato: Conceptualization, Methodology, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration

Haitham Ashoor: Conceptualization, Methodology, Formal analysis, Data Curation, Writing - Review & Editing, Visualization

Liping Hou: Methodology, Formal analysis, Writing - Review & Editing

Triantafillos Liloglou: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Funding acquisition

Rob Yang: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

John Field: Conceptualization, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

All authors have read and approved the final version of the manuscript. Michael Davies, Takahiro Sato and Haitham Ashoor verified the underlying data.

Data Sharing Statement

The data and materials supporting the conclusions of this article are included in this published article (and its supplementary information files). Primary Olink data and related sample metadata is available from the corresponding author on request, but individual level data can only be released under a suitable data sharing agreement due to informed consent restrictions.

Declaration of Interests

This work was funded through a Research Collaboration Agreement between Janssen Pharmaceuticals and the University of Liverpool: MD & JKF received research funding from Janssen Pharmaceuticals (a Johnson & Johnson company). TS, HA, LH & RY are employees of Johnson & Johnson, the company has filed a patent to on use of plasma protein biomarkers in lung cancer interception. TL declares no conflict of interest. Both parties shared responsibility for: study design; collection, analysis and interpretation of experimental data; writing the report and the decision to publish.

Acknowledgements

This work was supported through a Research Collaboration Agreement between Janssen Pharmaceuticals and the University of Liverpool. The Liverpool Lung Project has also been funded by the Roy Castle Lung Cancer Foundation, including a Senior Research fellowship for Dr Michael Davies. We would like to thank the Liverpool Lung Project participants for their invaluable contribution to this project and LLP staff past and present, especially Nial Hodge and Stephanie Tate for data and sample curation. We acknowledge staff at the Janssen World Without Disease Accelerator in Leiden who instigated this collaboration. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Table 1 LLP Cohorts used for 1-3 year and 1-5 year discovery

	Cases 1-3 years prior to diagnosis				Cases 1-5 years prior to diagnosis			
	Cancer	**Control**	**Total**	**P value (test)***	**Cancer**	**Control**	**Total**	**P value (test)***
Sex n (%) Female	14 (35.0)	39 (38.2)	53 (37.3)	X^2 0.13	27 (36.0)	77 (41.4)	104 (39.8)	X^2 0.65
Male	26 (65.0)	63 (61.8)	89 (62.7)	P= 0·72 (CS)	48 (64.0)	109 (58.6)	157 (60.2)	0·42 (CS)
Age (years)	69.5	70.1	69.8	0·96	68.3	68.2	68.1	0.88
Median (IQR)	(62.3 - 74.2)	(62.0 - 74.3)	(62.0 - 74.2)	(MW)	(62.0 - 73.3)	(61.9 - 73.2)	(62.0 - 73.2)	(MW)
Smoking status n (%) current	11 (27.5)	38 (37.3)	49 (34.5)	X^2 1.08	27 (36.0)	74 (39.8)	101 (38.7)	X^2 0.51
former	27 (67.5)	61 (59.8)	88 (62.0)	P= 0.58	43 (57.3)	104 (55.9)	147 (56.3)	P= 0·77
never	1 (2.5)	3 (2.9)	4 (2.8)	(CS)	2 (2.7)	8 (4.3)	10 (3.8)	(CS)
unknown	1 (2.5)	0 (0)	1 (0.7)		3 (4.0)	0 (0)	3 (1.1)	
Smoking duration (years)	44	43	43	0·47	44	44	44	0.76
Median (IQR)	(33 - 48)	(35 - 50)	(34 - 49)	(MW)	(34 - 49)	(35 - 49)	(35 - 49)	(MW)
Smoking pack years	43.5	39.8	39.9	0·68	41.3	37.5	38.4	0.19
Median (IQR)	(25.0 - 51.5)	(22.7 - 53.8)	(24.6 - 52.8)	(MW)	(25.5 - 51.8)	(21.8 - 49.2)	(23.3 - 50.4)	(MW)
Smoking quit years	0	2	0	0·75	0	0	0	0·59
Median (IQR)	(0 - 10)	(0 - 12.3)	(1 - 11.5)	(MW)	(0 - 10)	(0 - 9)	(0 - 8)	(MW)
COPD n (%) Yes	9 (22.5)	18 (17.6)	27 (19.0)	X^2 0.44	16 (21.3)	33 (17.7)	49 (18.8)	X^2 0.45
No	31 (77.5)	84 (82.4)	115 (81.0)	P= 0·51 (CS)	59 (78.7)	153 (82.3)	212 (81.2)	P= 0·50 (CS)
Body Mass Index	26.6	26.5	26.6	0·47	26.6	26.6	26.6	0.86
Median (IQR)	(26.2 - 29.3)	(24.3 - 28.1)	(24.6 - 28.2)	(MW)	(24.8 - 27.4)	(24.5 - 28.1)	(24.5 - 28.1)	(MW)
Total subjects	40	102	142		75	186	261	
Plasma samples	58	117	175		114	220	334	

IQR = Inter-quartile range; * CS = Chi-square; MW = Mann-Whitney (tests only performed for known values)

Table 2 Validation of 1-5Y lung cancer prediction model in UK Biobank data

	PPV at sensitivity of:			enrichment at 0·05	AUC	Population Size	Cases	Prevalence in subgroup
	0·05	0·10	0·25					
Smoker	47.4	37.1	21.7	5.6	0.693	4235	356	8.41
Non-smoker	7.7	8.1	6.6	3.9	0.615	1654	33	2
Age 40-55 y	100	62.5	27.9	39	0.775	1913	49	2.56
Age 55-70 y	30.4	31.5	21.3	3.5	0.683	3979	343	8.62
Male	55.6	29.9	20.2	7.8	0.721	2878	204	7.09
Female	31	31.7	17.6	5.0	0.663	3014	188	6.24
Total	40.8	30	19.1	6.1	0.694	5892	392	6.65

PPP = positive predictive value; AUC = Area under Curve ROC value

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021; **71**(3): 209-49.

2. Miller KD, Nogueira L, Mariotto AB, et al. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin* 2019; **69**(5): 363-85.

3. Nicholson AG, Chansky K, Crowley J, et al. The International Association for the Study of Lung Cancer Lung Cancer Staging Project: Proposals for the Revision of the Clinical and Pathologic Staging of Small Cell Lung Cancer in the Forthcoming Eighth Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* 2016; **11**(3): 300-11.

4. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med* 2020; **382**(6): 503-13.

5. Field JK, Vulkan D, Davies MPA, et al. Lung cancer mortality reduction by LDCT screening: UKLS randomised trial results and international meta-analysis. *Lancet Reg Health Eur* 2021; **10**: 100179.

6. National Lung Screening Trial Research T, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011; **365**(5): 395-409.

7. Ten Haaf K, van der Aalst CM, de Koning HJ, Kaaks R, Tammemagi MC. Personalising lung cancer screening: An overview of risk-stratification opportunities and challenges. *Int J Cancer* 2021; **149**(2): 250-63.

8. Seijo LM, Peled N, Ajona D, et al. Biomarkers in Lung Cancer Screening: Achievements, Promises, and Challenges. *J Thorac Oncol* 2019; **14**(3): 343-57.

9. Dama E, Colangelo T, Fina E, et al. Biomarkers and Lung Cancer Early Detection: State of the Art. *Cancers (Basel)* 2021; **13**(15).

10. Ostrin EJ, Sidransky D, Spira A, Hanash SM. Biomarkers for Lung Cancer Screening and Detection. *Cancer Epidemiol Biomarkers Prev* 2020; **29**(12): 2411-5.

11. Fahrmann JF, Marsh T, Irajizad E, et al. Blood-Based Biomarker Panel for Personalized Lung Cancer Risk Assessment. *J Clin Oncol* 2022; **40**(8): 876-83.

12. INTEGRAL Consortium for Early Detection of Lung Cancer, Guida F, Sun N, et al. Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins. *JAMA Oncol* 2018; **4**(10): e182078.

13. Dagnino S, Bodinier B, Guida F, et al. Prospective Identification of Elevated Circulating CDCP1 in Patients Years before Onset of Lung Cancer. *Cancer Res* 2021; **81**(13): 3738-48.

14. Field JK, Smith DL, Duffy S, Cassidy A. The Liverpool Lung Project research protocol. *Int J Oncol* 2005; **27**(6): 1633-45.

15. Olink. Pre-analytical variation in protein biomarker research. 13/5/2020 2020. https://olink.com/application/variation-in-biomarker-research/ (accessed 19/11/2022 2022).

16. Sun BB, Chiou J, Traylor M, et al. Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* 2022.

17. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005; **67**(2): 301-20.

18. Tin Kam H. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995 14-16 Aug. 1995; 1995. p. 278-82 vol.1.

19. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995; **20**(3): 273-97.

20. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Kdd '16* 2016: 785–94.

21. Ridker PM, MacFadyen JG, Thuren T, et al. Effect of interleukin-1beta inhibition with canakinumab on incident lung cancer in patients with atherosclerosis: exploratory results from a randomised, double-blind, placebo-controlled trial. *Lancet* 2017; **390**(10105): 1833-42.

22. Klein EA, Richards D, Cohn A, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol* 2021; **32**(9): 1167-77.

23. Mathios D, Johansen JS, Cristiano S, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* 2021; **12**(1): 5060.

24. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018; **359**(6378): 926-30.

25. Abbosh C, Birkbak NJ, Wilson GA, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* 2017; **545**(7655): 446-51.

26. Ooki A, Maleki Z, Tsay JJ, et al. A Panel of Novel Detection and Prognostic Methylated DNA Markers in Primary Non-Small Cell Lung Cancer and Serum DNA. *Clin Cancer Res* 2017; **23**(22): 7141-52.

27. Integrative Analysis of Lung Cancer E, Risk Consortium for Early Detection of Lung C, Guida F, et al. Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins. *JAMA Oncol* 2018; **4**(10): e182078.

28. Wang G, Qiu M, Xing X, et al. Lung cancer scRNA-seq and lipidomics reveal aberrant lipid metabolism for early-stage diagnosis. *Sci Transl Med* 2022; **14**(630): eabk2756.

29. Montani F, Marzi MJ, Dezi F, et al. miR-Test: a blood test for lung cancer early detection. *J Natl Cancer Inst* 2015; **107**(6): djv063.

30. Sozzi G, Boeri M, Rossi M, et al. Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. *J Clin Oncol* 2014; **32**(8): 768-73.

31. Pastorino U, Boeri M, Sestini S, et al. Baseline computed tomography screening and blood microRNA predict lung cancer risk and define adequate intervals in the BioMILD trial. *Ann Oncol* 2022; **33**(4): 395-405.

32. Robbins HA, Alcala K, Moez EK, et al. Design and methodological considerations for biomarker discovery and validation in the Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Program. *Ann Epidemiol* 2023; **77**: 1-12.

Figure 1: Circulating plasma proteins prediction of future lung cancer. (a) A boxplot of training AUC values from four different machine learning models (Elastic Net, Random Forest, Support Vector Machine, XGBoost, 5-fold CV repeated 5 times) trained on the LLP cohort to predict lung cancer in patients 1-3 years before diagnosis (53 cancer and 109 control samples). (b) Protein levels in LLP subjects were transformed using the z-score method and combined to generate one score. Combined z-scores were plotted over time in the LLP cohort for 1-3Y proteins. (c) AUROC of 1-3Y SVM model trained in Liverpool tested in UK Biobank samples 1-3 years before lung cancer diagnosis (62 cancer and 5500 control samples). (d) Performance of the 1-3Y SVM model in the UK Biobank across different years of diagnosis of lung cancer. Samples taken at different times prior to lung cancer were segregated by year (2-12 years) and the SVM model for 1-3Y was tested by ROC analysis. (e) Barplot for AUROC values for SVM model predicting future development of cancer for several cancer types from UK Biobank 1-3 years before diagnosis. The same approach as taken for lung cancer (see methods) was taken to identify plasma samples at least 2 years prior to other first cancer diagnosis (number of cases labelled on bar chart) and the AUC for ROC analysis shown.

Figure 2: Combined z-score from 1-3Y in relation to cancer stage and pack years of smoking. Protein levels in LLP subjects were transformed using the z-score method and combined to generate one score. (a) Combined z-scores were plotted in time-frame categories (5-10 years, 3-5 years, 1-3 years prior to diagnosis or at diagnosis) for healthy subjects and cases of different lung cancer stage for 1-3Y proteins with P-values generated using Wilcoxon signed-rank test. (b) The z-scores were also correlated with pack years of smoking at time of sample in the same time frame categories; correlation was measured using Pearson correlation coefficient.

Figure 3: Circulating plasma proteins prediction of long-term future lung cancer. (a) A boxplot of training AUC values from four different machine learning models (Elastic Net, Random Forest, Support Vector Machine, XGBoost, 5-fold CV repeated 5 times) trained on the LLP cohort to predict lung cancer in patients 1-5 years before diagnosis (110 Cancer, 215 control samples). (b) Protein levels in LLP subjects were transformed using the z-score method and combined to generate one score. Combined z-scores were plotted over time in the LLP cohort for 1-5Y proteins. (c) The z-scores were also correlated with age at time of sample in the same time frame categories; correlation was measured using Pearson correlation coefficient.

Figure 4: Gene Enrichment Analysis Top 20 pathways over- or under-represented in plasma samples from 1-3Y or 1-5Y models, demonstrating largely different pathways for different predictive panels (blue) with three shared over-represented (green) and three shared under-represented (red) pathways.

Figure 1

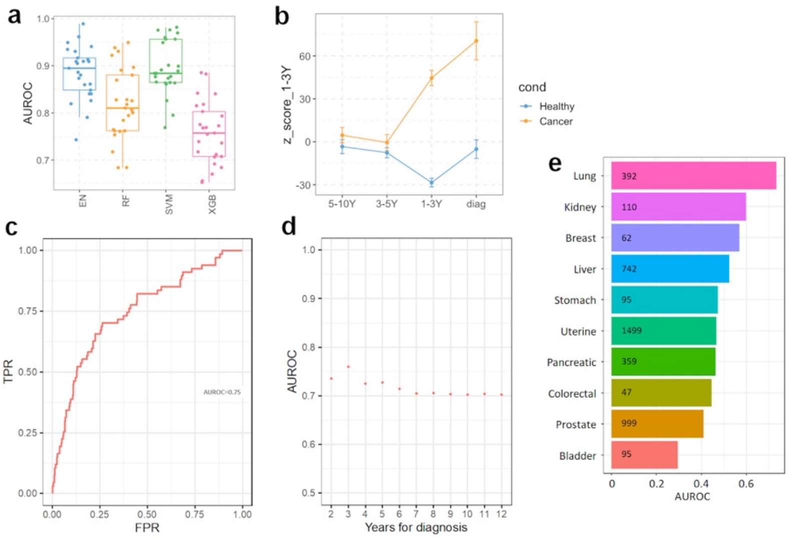

a
AUROC
1.0
0.9
0.8
0.7
EN
RF
SVM
XGB
b
z_score_1-3Y
60
30
0
-30
5-10Y
3-5Y
1-3Y
diag
cond
Healthy
Cancer
c
TPR
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
FPR
AUROC=0.75
d
AUROC
1.0
0.9
0.8
0.7
0.6
0.5
2
3
4
5
6
7
8
9
10
11
12
Years for diagnosis
e
Lung
392
Kidney
110
Breast
62
Liver
742
Stomach
95
Uterine
1499
Pancreatic
359
Colorectal
47
Prostate
999
Bladder
95
0
0.2
0.4
0.6
AUROC

Figure 2

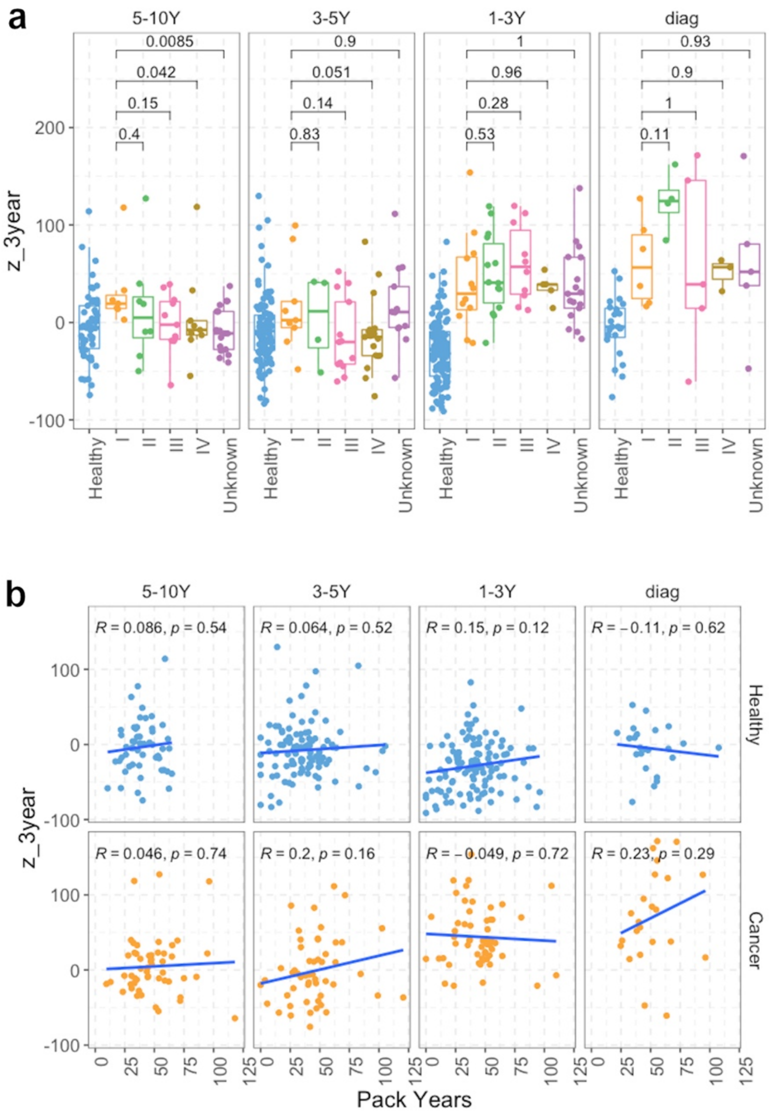

a
5-10Y
3-5Y
1-3Y
diag
0.0085
0.042
0.15
0.4
0.9
0.051
0.14
0.83
1
0.96
0.28
0.53
0.93
0.9
1
0.11
z_3year
200
100
0
-100
Healthy
I
II
III
IV
Unknown
b
R = 0.086, p = 0.54
R = 0.064, p = 0.52
R = 0.15, p = 0.12
R = −0.11, p = 0.62
R = 0.046, p = 0.74
R = 0.2, p = 0.16
R = −0.049, p = 0.72
R = 0.23, p = 0.29
Healthy
Cancer
0
25
50
75
100
125
Pack Years

Figure 3

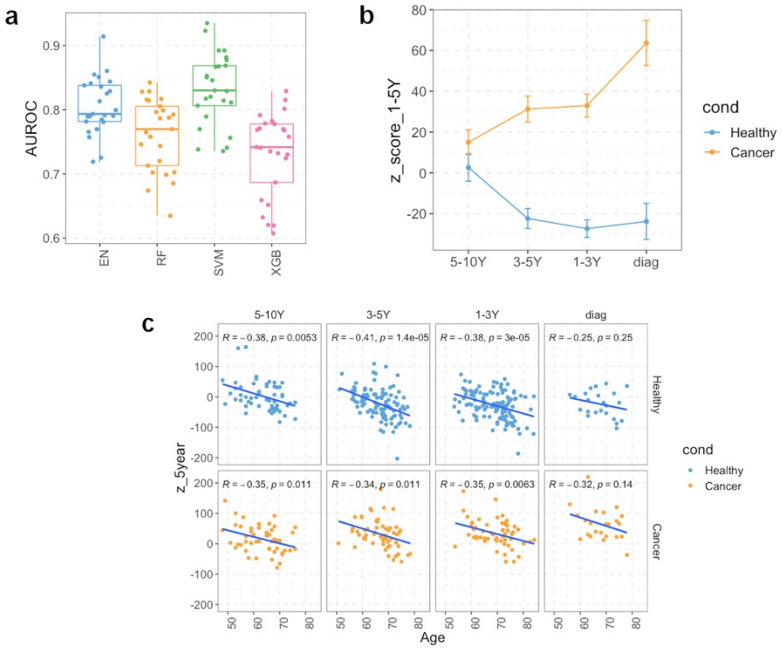
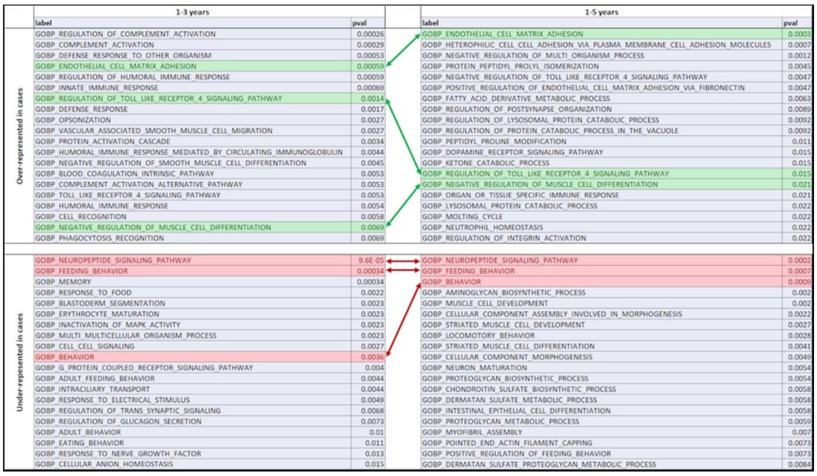

a
AUROC
0.9
0.8
0.7
0.6
EN
RF
SVM
XGB
b
z_score_1-5Y
80
60
40
20
0
-20
5-10Y
3-5Y
1-3Y
diag
cond
Healthy
Cancer
c
5-10Y
3-5Y
1-3Y
diag
R = −0.38, p = 0.0053
R = −0.41, p = 1.4e-05
R = −0.38, p = 3e-05
R = −0.25, p = 0.25
R = −0.35, p = 0.011
R = −0.34, p = 0.011
R = −0.35, p = 0.0063
R = −0.32, p = 0.14
Healthy
Cancer
z_5year
200
100
0
-100
-200
50
60
70
80
Age
cond
Healthy
Cancer

Figure 4



	1-3 years			1-5 years	
	label	pval		label	pval
Over-represented in cases	GOBP_REGULATION_OF_COMPLEMENT_ACTIVATION	0.00026		GOBP_ENDOTHELIAL_CELL_MATRIX_ADHESION	0.0003
	GOBP_COMPLEMENT_ACTIVATION	0.00029		GOBP_HETEROPHILIC_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_CELL_ADHESION_MOLECULES	0.0007
	GOBP_DEFENSE_RESPONSE_TO_OTHER_ORGANISM	0.00053		GOBP_NEGATIVE_REGULATION_OF_MULTI_ORGANISM_PROCESS	0.0012
	GOBP_ENDOTHELIAL_CELL_MATRIX_ADHESION	0.00059		GOBP_PROTEIN_PEPTIDYL_PROLYL_ISOMERIZATION	0.0045
	GOBP_REGULATION_OF_HUMORAL_IMMUNE_RESPONSE	0.00059		GOBP_NEGATIVE_REGULATION_OF_TOLL_LIKE_RECEPTOR_4_SIGNALING_PATHWAY	0.0047
	GOBP_INNATE_IMMUNE_RESPONSE	0.00069		GOBP_POSITIVE_REGULATION_OF_ENDOTHELIAL_CELL_MATRIX_ADHESION_VIA_FIBRONECTIN	0.0047
	GOBP_REGULATION_OF_TOLL_LIKE_RECEPTOR_4_SIGNALING_PATHWAY	0.0014		GOBP_FATTY_ACID_DERIVATIVE_METABOLIC_PROCESS	0.0063
	GOBP_DEFENSE_RESPONSE	0.0017		GOBP_REGULATION_OF_POSTSYNAPSE_ORGANIZATION	0.0089
	GOBP_OPSONIZATION	0.0027		GOBP_REGULATION_OF_LYSOSOMAL_PROTEIN_CATABOLIC_PROCESS	0.0092
	GOBP_VASCULAR_ASSOCIATED_SMOOTH_MUSCLE_CELL_MIGRATION	0.0027		GOBP_REGULATION_OF_PROTEIN_CATABOLIC_PROCESS_IN_THE_VACUOLE	0.0092
	GOBP_PROTEIN_ACTIVATION_CASCADE	0.0034		GOBP_PEPTIDYL_PROLINE_MODIFICATION	0.011
	GOBP_HUMORAL_IMMUNE_RESPONSE_MEDIATED_BY_CIRCULATING_IMMUNOGLOBULIN	0.0044		GOBP_DOPAMINE_RECEPTOR_SIGNALING_PATHWAY	0.015
	GOBP_NEGATIVE_REGULATION_OF_SMOOTH_MUSCLE_CELL_DIFFERENTIATION	0.0045		GOBP_KETONE_CATABOLIC_PROCESS	0.015
	GOBP_BLOOD_COAGULATION_INTRINSIC_PATHWAY	0.0053		GOBP_REGULATION_OF_TOLL_LIKE_RECEPTOR_4_SIGNALING_PATHWAY	0.015
	GOBP_COMPLEMENT_ACTIVATION_ALTERNATIVE_PATHWAY	0.0053		GOBP_NEGATIVE_REGULATION_OF_MUSCLE_CELL_DIFFERENTIATION	0.021
	GOBP_TOLL_LIKE_RECEPTOR_4_SIGNALING_PATHWAY	0.0053		GOBP_ORGAN_OR_TISSUE_SPECIFIC_IMMUNE_RESPONSE	0.021
	GOBP_HUMORAL_IMMUNE_RESPONSE	0.0054		GOBP_LYSOSOMAL_PROTEIN_CATABOLIC_PROCESS	0.022
	GOBP_CELL_RECOGNITION	0.0058		GOBP_MOLTING_CYCLE	0.022
	GOBP_NEGATIVE_REGULATION_OF_MUSCLE_CELL_DIFFERENTIATION	0.0069		GOBP_NEUTROPHIL_HOMEOSTASIS	0.022
	GOBP_PHAGOCYTOSIS_RECOGNITION	0.0069		GOBP_REGULATION_OF_INTEGRIN_ACTIVATION	0.022
Under-repesented in cases	GOBP_NEUROPEPTIDE_SIGNALING_PATHWAY	9.6E-05		GOBP_NEUROPEPTIDE_SIGNALING_PATHWAY	0.0002
	GOBP_FEEDING_BEHAVIOR	0.00034		GOBP_FEEDING_BEHAVIOR	0.0007
	GOBP_MEMORY	0.00034		GOBP_BEHAVIOR	0.0009
	GOBP_RESPONSE_TO_FOOD	0.0022		GOBP_AMINOGLYCAN_BIOSYNTHETIC_PROCESS	0.002
	GOBP_BLASTODERM_SEGMENTATION	0.0023		GOBP_MUSCLE_CELL_DEVELOPMENT	0.002
	GOBP_ERYTHROCYTE_MATURATION	0.0023		GOBP_CELLULAR_COMPONENT_ASSEMBLY_INVOLVED_IN_MORPHOGENESIS	0.0022
	GOBP_INACTIVATION_OF_MAPK_ACTIVITY	0.0023		GOBP_STRIATED_MUSCLE_CELL_DEVELOPMENT	0.0027
	GOBP_MULTI_MULTICELLULAR_ORGANISM_PROCESS	0.0023		GOBP_LOCOMOTORY_BEHAVIOR	0.0028
	GOBP_CELL_CELL_SIGNALING	0.0027		GOBP_STRIATED_MUSCLE_CELL_DIFFERENTIATION	0.0041
	GOBP_BEHAVIOR	0.0036		GOBP_CELLULAR_COMPONENT_MORPHOGENESIS	0.0049
	GOBP_G_PROTEIN_COUPLED_RECEPTOR_SIGNALING_PATHWAY	0.004		GOBP_NEURON_MATURATION	0.0054
	GOBP_ADULT_FEEDING_BEHAVIOR	0.0044		GOBP_PROTEOGLYCAN_BIOSYNTHETIC_PROCESS	0.0054
	GOBP_INTRACILIARY_TRANSPORT	0.0044		GOBP_CHONDROITIN_SULFATE_BIOSYNTHETIC_PROCESS	0.0058
	GOBP_RESPONSE_TO_ELECTRICAL_STIMULUS	0.0049		GOBP_DERMATAN_SULFATE_METABOLIC_PROCESS	0.0058
	GOBP_REGULATION_OF_TRANS_SYNAPTIC_SIGNALING	0.0068		GOBP_INTESTINAL_EPITHELIAL_CELL_DIFFERENTIATION	0.0058
	GOBP_REGULATION_OF_GLUCAGON_SECRETION	0.0073		GOBP_PROTEOGLYCAN_METABOLIC_PROCESS	0.0059
	GOBP_ADULT_BEHAVIOR	0.01		GOBP_MYOFIBRIL_ASSEMBLY	0.007
	GOBP_EATING_BEHAVIOR	0.011		GOBP_POINTED_END_ACTIN_FILAMENT_CAPPING	0.0073
	GOBP_RESPONSE_TO_NERVE_GROWTH_FACTOR	0.013		GOBP_POSITIVE_REGULATION_OF_FEEDING_BEHAVIOR	0.0073
	GOBP_CELLULAR_ANION_HOMEOSTASIS	0.015		GOBP_DERMATAN_SULFATE_PROTEOGLYCAN_METABOLIC_PROCESS	0.0084