**Assessment of lung cancer risk based on a biomarker panel of circulating proteins**

**Front page group name:** The INTEGRAL consortium for early detection of lung cancer

**Full author list (PubMed searchable) to be listed in the Group information section:**

Florence Guida, PhD1\*, Nan Sun, PhD2\*, Leonidas E. Bantis, PhD2\*, David C. Muller, PhD4\*, Peng Li, PhD1,3, Ayumu Taguchi, MD, PhD2, Dilsher Dhillon, MS2, Deepali L. Kundnani, MS2, Nikul J. Patel, MS2, Qingxiang Yan, PhD2, Graham Byrnes, PhD1, Karel G M Moons, PhD5, Anne Tjønneland, MD, PhD6, Salvatore Panico, MD, MS7, Claudia Agnoli, PhD8, Paolo Vineis, MD, MPH, FFPH4,9, Domenico Palli, MD10, Bas Bueno-de-Mesquita, MD, MPH,PhD4,11, Petra H. Peeters, MD, PhD5,  Antonio Agudo, MD, PhD12,  Jose M. Huerta, PhD13,14,  Miren Dorronsoro, MD15,  Miguel Rodriguez Barranco, PhD14,16,17,  Eva Ardanaz, MD, PhD14,18,19,  Ruth C. Travis, DPhil20,  Karl Smith Byrne, DPhil20,  Heiner Boeing, PhD21,  Annika Steffen, PhD21,  Rudolf Kaaks, PhD22,  Anika Hüsing, MS22,  Antonia Trichopoulou, PhD23,24,  Pagona Lagiou, MD23,24,25,  Carlo La Vecchia, MD23,26,  Gianluca Severi, PhD9,27,  Marie-Christine Boutron-Ruault, PhD27,  Torkjel M. Sandanger, PhD28,  Elisabete Weiderpass Vainio, MD, PhD28,29,30,31,  Therese H. Nøst, PhD28,  Kostas Tsilidis, PhD4,32,  Elio Riboli, MD, MPH, MSc4,  Kjell Grankvist, MD, PhD33,  Mikael Johansson, PhD33,  Gary E. Goodman, MD, MS34,  Ziding Feng, PhD2, Paul Brennan, PhD1, Mattias Johansson, PhD1\*\*, Samir M. Hanash, MD, PhD2\*\*.

\*These authors contributed equally to this study

\*\*These authors supervised the study

**Affiliation/Institution:**

1, International Agency for Research on Cancer, Lyon, France

2, University of Texas MD Anderson Cancer Center, Houston, TX, USA

3, Max Planck Institute for Demographic Research, Rostock, Germany

4, Imperial College London School of Public Health, London, UK

5, Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, Netherlands

6, Danish Cancer Society Research Center, Copenhagen, Denmark

7, Federico II University, Naples, Italy

8, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

9, Human Genetics Foundation, Torino, Italy

10, Cancer Research and Prevention Institute–ISPO, Florence, Italy

11, National Institute for Public Health and the Environment, Bilthoven, Netherlands

12, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

13, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain

14, CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

15, Public Health Direction and Biodonostia Research Institute-CIBERESP, San Sebastian, Spain

16, Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria ibs, Granada, Spain

17, Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain

18, Navarra Public Health Institute, Pamplona, Spain

19, IdiSNA, Navarra Institute for Health Research, Pamplona, Spain

20, Nuffield Department of Population Health, University of Oxford, Oxford, UK

21, German Institute of Human Nutrition, Potsdam-Rehbruecke, Germany

22, DKFZ-German Cancer Research Center, Heidelberg, Germany

23, Hellenic Health Foundation, Athens, Greece

24, WHO Collaborating Center for Nutrition and Health, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece

25, Harvard School of Public Health, Boston, USA

26, Università degli Studi di Milano, Milano, Italy

27, Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM, Villejuif, France

28, UiT-The Arctic University of Norway, Tromsø, Norway

29, Institute of Population-Based Cancer Research, Oslo, Norway

30, Karolinska Institutet, Stockholm, Sweden

31, Folkhälsan Research Center, Helsinki, Finland

32, School of Medicine, University of Ioannina, Ioannina, Greece

33, Umeå University, Umeå, Sweden

34, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

**Contact information for the corresponding authors:**

Dr. Samir M. Hanash:

University of Texas MD Anderson Cancer Center, 6767 Bertner Ave, Houston, TX 77030, USA.

Tel: +1-713-745-5242,

Fax: +1-713-792-1474,

E-mail: SHanash@mdanderson.org

Dr. Mattias Johansson:

International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon CEDEX 08, France.

Tel: +33-472738023

Fax: +33-472738320

E-mail: JohanssonM@iarc.fr

Word count of the manuscript text: 1,231 words

Date of the revision: 23rd March 2018

**Key points:**

**Question:** Can a panel of circulating protein biomarkers improve upon a traditional risk prediction model for lung cancer?

**Findings:** Abiomarker-based risk-prediction model consisting of four protein markers that was developed in a high-risk US cohort outperformed a model based on smoking history alone when blindly validated using pre-diagnostic samples from two European cohorts.

**Meaning:** Biomarker-based risk profiling has the potential to improve eligibility criteria for screening.

**Importance:**  There is an urgent need to improve lung cancer risk assessment as current screening criteria miss a large proportion of cases.

**Objective:** To determine if a panel of selected circulating protein biomarkers can contribute to lung cancer risk assessment and outperform current US screening criteria.

**Design, Setting and Participants:**  Pre-diagnostic samples from ever-smoking cases diagnosed within one year of blood collection and smoking-matched controls from the Carotene and Retinol Efficacy Trial (CARET) cohort were used to develop a biomarker risk-score based on 4 proteins (CA125, CEA, CYFRA 21-1 and Pro-SFTPB). The biomarker score was subsequently validated blindly using absolute risk-estimates in ever-smoking cases diagnosed within one year of blood collection and matched controls from two large European population-based cohorts; the European Prospective Investigation into Cancer and nutrition (EPIC) study and the Northern Sweden Health and Disease Study (NSHDS).

**Main Outcome and Measures**: Model validity in discriminating between future lung cancer cases and controls. Discrimination estimates were weighted to reflect the background populations of EPIC and NSHDS validation studies (area under receiver-operating characteristics curve [AUC], sensitivity and specificity).

**Results:** In the validation study, an integrated risk-prediction model combining smoking exposure with the biomarker score yielded an AUC of 0.83 (95% CI: 0.76-0.90) compared to 0.73 (95% CI: 0.64-0.82) for a model based on smoking exposure alone (P=0.003 for difference in AUC). At an overall specificity of 0.83 based on the USPSTF screening criteria, the sensitivity of the integrated risk-prediction model (biomarker) model was 0.63 compared to 0.43 for the smoking model. Conversely, at an overall sensitivity of 0.42 (USPSTF), the integrated risk-prediction model yielded a specificity of 0.95 compared to 0.86 for the smoking model.

**Conclusions and Relevance:** This study provided a proof-of-principle in demonstrating that a panel of circulating protein biomarkers can improve lung cancer risk assessment and may be used to define eligibility for CT-screening.

**Funding:** US NCI/NIH, Fondation ARC, INCa, MD Anderson Lung Cancer Moon Shot Program, Lyda Hill Foundation, Canary Foundation, Rubenstein Family Foundation.

The National Lung Screening Trial (NLST) findings suggested that screening with low-dose computed tomography (LDCT) can reduce lung cancer mortality.1 As a result, the US Preventive Services Task Force (USPSTF) recommends LDCT screening for lung cancer among individuals aged 55-80 years who have smoked 30 pack-years with up to 15 years since quitting smoking.1,2 However, LDCT screening results in a large number of indeterminate nodules,1 and less than 50% of incident lung cancer cases are eligible for screening.3 Biomarkers may improve lung cancer risk assessment over and beyond that of traditional smoking-based risk models and improve current screening eligibility criteria.4,5

We demonstrated that the precursor form of surfactant protein B (Pro-SFTPB) is predictive of lung cancer risk.5,6 Other markers that have been shown to be useful for the work-up and diagnosis of lung cancer include cancer antigen 125 (CA125), cytokeratin-19 fragment (CYFRA 21-1), carcinoembryonic antigen (CEA), and Human Epididymis Protein 4 (HE4).7-12 However, there is limited data regarding their performance in discriminating between future lung cancer cases and controls.

The current study aimed to assess the potential of these 5 protein biomarkers to inform about lung cancer risk when tested blindly using pre-diagnostic samples.

**Methods**

A full account of the methods are provided in the Supplement.

Briefly, samples from ever-smoking lung cancer cases diagnosed within one year following blood collection (N=108) and smoking-matched controls (N=216) from the US CARET cohort were used to develop a biomarker score based on circulating measures of Pro-SFTPB, CA125, CEA, HE4 and CYFRA 21-1 using logistic regression.

The extent to which the biomarker score improved discrimination of incident lung cancer cases and controls was validated externally using ever-smoking lung cancer cases diagnosed within one year following blood collection (N=63) and matched controls (N=90) from the European EPIC and NSHDS cohorts (eFigure 1). Absolute 1-year risks of lung cancer were estimated for each study participant in the validation study by modelling the cumulative hazards of lung cancer using flexible parametric survival models.13 Two models were evaluated, *i)* a traditional smoking history-based risk model, and *ii)* an integrated risk-prediction model combining the smoking model and the biomarker score. Model discrimination was assessed by receiver operating characteristics (ROC) analysis using the predicted 1-year lung cancer risks as scoring rule. Discrimination estimates included area under ROC curve (AUC), sensitivity and specificity that were weighted to reflect the background populations.

Statistical significance was assumed at a two-sided P-value below 0.05. All study participants gave written informed consent to participate in the study and the research was approved by the local ethics committees.

**Results**

Details of the biomarker score and discrimination estimates in the CARET training study are available in eTable 1 and 2, eFigure 2 and 3.

In the validation study, the predicted 1-year risk for a 60-year old male with 30 pack-years of smoking history was estimated at 0.37% using the smoking model (Figure 1). In comparison, using the integrated risk-prediction model we estimated 1-year risks at 0.07% and 1.56% for the same man assuming a biomarker score in the 1st and 4th quartile, respectively. The 1-year lung cancer risk estimates for each study participant in the validation study according to the smoking and integrated risk-prediction models are depicted in Figure 2. In comparison to the smoking model, the median 1-year risk estimates from the integrated risk-prediction model increased for cases from 0.27% to 0.45%, and decreased for controls 0.12% to 0.04%.

In the validation study, the population-weighted AUC was 0.73 (95% CI: 0.64-0.82) for the smoking model and 0.83 (95% CI: 0.76-0.90) for the integrated risk-prediction model (P=0.003 for difference in AUC, Figure 3-A). The AUCs were consistently higher for the integrated model than for the smoking model across relevant strata (eTable 3). At an overall specificity of 0.83 based on the USPSTF screening criteria, the integrated risk-prediction model yielded a sensitivity of 0.63 (95% CI: 0.49-0.76) compared to 0.43 (95% CI: 0.23-0.65) for the smoking model. Similarly, at an overall sensitivity of 0.42 (USPSTF), the integrated risk-prediction model yielded a specificity of 0.95 (95% CI: 0.85-0.99) compared to 0.86 (95% CI: 0.72-0.94) for the smoking model. The improvement in AUC for the integrated risk prediction model (AUC: 0.80, 95% CI: 0.75-0.85) over the smoking model (AUC: 0.73, 95% CI: 0.68-0.79) was more modest when considering cases diagnosed up to 2 years after blood draw (eFigure 4).

A full account of all conducted analyses are provided in the Supplement.

**Discussion**

This is the first study in which a blood-based biomarker score was developed using one cohort and externally validated using pre-diagnostic samples from other independent cohorts. We observed a notable improvement in discrimination between future lung cancer cases and controls over a traditional smoking-based risk-prediction model by incorporating information from a biomarker score consisting of four circulating proteins.

In our validation study, 42% of the incident lung cancer cases would have qualified for CT screening according to USPSTF criteria. Using the biomarker score together with smoking information, we estimated that 63% of *all* cases could be identified, without increasing the number of false positives. The data further suggested that the biomarker score could alternatively be used to reduce screening of subjects not destined to develop lung cancer from 14% to 5%, without affecting the uptake of future lung cancer cases. These improvements in sensitivity and specificity were consistently observed across each evaluated stratum. Our findings also indicated that the improvement in discrimination afforded by the biomarker score is more modest beyond the initial year after blood draw, suggesting that an annual biomarker test may be necessary in a a screening program.

**Strengths and Limitations**

Naïve discrimination estimates, as typically provided in a matched nested case-control setting, are inherently biased. An important strength of our study was the use of absolute risks and population-based discrimination estimates, which are necessary to evaluate the impact of applying a biomarker tool in a practical screening scenario.

A limitation of our study was that three variables originally included in a validated risk prediction model (PLCOM2012) were not available in our validation studies. However, using the original PLCO data, excluding these variables from the PLCOM2012 model only nominally decreased its performance, suggesting that our risk prediction model represented a valid comparison for the biomarkers score (eMethods, eFigure 5).14

Whilst the current study provided a proof-of-principle of the potential of using biomarkers in lung cancer risk assessment to define screening eligibility, validating and calibrating the integrated risk-prediction model using larger sample size with pre-diagnostic samples is clearly needed before such a risk prediction tool can be used in practice. A larger sample size will also allow stratified analysis aiming to evaluate the performance of the biomarker panel in predicting lung cancer cases of different characteristics, in particular by stage at diagnosis and histological subtype. Further, our study was limited to a select panel of circulating proteins, and we note that other types of biomarkers may also be informative.4,5 We also note that the population that would most benefit from a biomarker test prior to undergoing LDCT screening remains to be defined. A thorough cost-effectiveness assessment based on a large study sample is warranted to determine the threshold in absolute risk for developing lung cancer over a specific time-period, above which the benefits of screening outweigh the harms.15

**Conclusion**

The current study provided a proof-of-principle in demonstrating that circulating biomarkers have the potential to inform lung cancer risk assessment and substantially improve upon current criteria for CT-screening.**Acknowledgment**

*Contributions*

SH, MJ, and PBr initiated, acquired the main funding, and designed this investigation. SH,NS, AT, DD, DK, and NP led the laboratory analysis. FG and DM conducted the statistical analysis of EPIC and NSHDS data under the supervision of MJ. LB and QY conducted the statistical analysis of the CARET data under supervision of ZF. MJ, PBr, NS, LB, FG, ZF and SH drafted the first version of the manuscript. All authors were involved with collection of data, data interpretation, critical revisions of the paper, and approval of the final version. SH had had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. ER is the overall coordinator of the EPIC study, which he designed and implemented in collaboration with the main investigators in the collaborating centers.

*Conflicts of Interest and Financial Disclosures***:** The authors declare no conflict of interest.

*Funding*

This study was supported by the US National Cancer Institute (1U19CA203654, UO1194733, NCI Early Detection Research Network), Fondation ARC pour la recherche sur le cancer and INCa (INCa\_ARC\_10450), the MD Anderson Lung Cancer Moon Shot Program and the Lyda Hill Foundation, The Canary Foundation, The Lungevity Foundation, and the S. Rubenstein Family Foundation.

The funding organizations had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The EPIC study has been supported by the Europe Against Cancer Program of the European Commission (SANCO); Deutsche Krebshilfe; Deutsches Krebsforschungszentrum; German Federal Ministry of Education and Research; Danish Cancer Society; Health Research Fund (FIS) of the Spanish Ministry of Health; Spanish Regional Governments of Andalucia, Asturias, Basque Country, Murcia and Navarra; Catalan Institute of Oncology, Spain; the ISCIII of the Spanish Ministry of Health (RETICC DR06/0020); Cancer Research UK; Medical Research Council, United Kingdom; Greek Ministry of Health; Stavros Niarchos Foundation; Hellenic Health Foundation; Italian Association for Research on Cancer (AIRC); Italian National Research Council; Fondazione-Istituto Banco Napoli, Italy; Associazione Italiana per la Ricerca sul Cancro-AIRC-Milan; Compagnia di San Paolo; Dutch Ministry of Public Health, Welfare and Sports; World Cancer Research Fund; Swedish Cancer Society; Swedish Scientific Council; Regional Government of Västerbotten, Sweden; NordForsk (Centre of excellence programme HELGA), Norway; French League against Cancer (LNCC), France; National Institute for Health and Medical Research (INSERM), France; Mutuelle Générale de l’Education Nationale (MGEN), France; 3M Co, France; Gustave Roussy Institute (IGR), France; and General Councils of France.

**References**

1. 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.

2. Humphrey L, Deffebach M, Pappas M, et al. In: Screening for Lung Cancer: Systematic Review to Update the U.S. Preventive Services Task Force Recommendation. Rockville (MD)2013.

3. Muller DC, Johansson M, Brennan P. Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study. J Clin Oncol. 2017;35(8):861-869.

4. Shiels MS, Pfeiffer RM, Hildesheim A, et al. Circulating Inflammation Markers and Prospective Risk for Lung Cancer. Jnci-Journal of the National Cancer Institute. 2013;105(24):1871-1880.

5. Sin DD, Tammemagi CM, Lam S, et al. Pro–Surfactant Protein B as a Biomarker for Lung Cancer Prediction. Journal of Clinical Oncology. 2013;31(36):4536-4543.

6. Taguchi A, Hanash S, Rundle A, et al. Circulating Pro-Surfactant Protein B as a Risk Biomarker for Lung Cancer. Cancer Epidemiology Biomarkers & Prevention. 2013;22(10):1756-1761.

7. Bigbee WL, Gopalakrishnan V, Weissfeld JL, et al. A Multiplexed Serum Biomarker Immunoassay Panel Discriminates Clinical Lung Cancer Patients from High-Risk Individuals Found to be Cancer-Free by CT Screening. Journal of Thoracic Oncology. 2012;7(4):698-708.

8. Rastel D, Ramaioli A, Cornillie F, Thirion B. CYFRA 21-1, a sensitive and specific new tumour marker for squamous cell lung cancer. Report of the first European multicentre evaluation. CYFRA 21-1 Multicentre Study Group. Eur J Cancer. 1994;30A(5):601-606.

9. Patz EF, Jr., Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, Herndon JE, 2nd. Panel of serum biomarkers for the diagnosis of lung cancer. J Clin Oncol. 2007;25(35):5578-5583.

10. Schneider J, Bitterlich N, Kotschy-Lang N, Raab W, Woitowitz HJ. A fuzzy-classifier using a marker panel for the detection of lung cancers in asbestosis patients. Anticancer Research. 2007;27(4A):1869-1877.

11. Zeng Q, Liu M, Zhou N, Liu L, Song X. Serum human epididymis protein 4 (HE4) may be a better tumor marker in early lung cancer. Clin Chim Acta. 2016;455:102-106.

12. Taguchi A, Politi K, Pitteri SJ, et al. Lung Cancer Signatures in Plasma Based on Proteome Profiling of Mouse Tumor Models. Cancer Cell. 2011;20(3):289-299.

13. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175-2197.

14. Tammemagi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. N Engl J Med. 2013;368(8):728-736.

15. Ten Haaf K, van Rosmalen J, de Koning HJ. Lung cancer detectability by test, histology, stage, and gender: estimates from the NLST and the PLCO trials. Cancer Epidemiol Biomarkers Prev. 2015;24(1):154-161.

**Figure Legend**

**Figure 1**. 1-year predicted lung cancer probability for a male from NSHDS according to his smoking history.

A: predicted lung cancer probability from the smoking risk prediction model based on age (in years) and smoking history. The “rug plot” shows the observed distribution of age in the validation study (EPIC and NSHDS, ever smokers). B: predicted lung cancer probability from the integrated risk prediction model according to the biomarker score and the smoking history. The “rug plot” shows the observed distribution of the biomarker score in the validation study (EPIC and NSHDS, ever smokers). The vertical lines correspond to the quartiles threshold for biomarker score among controls (Q1, Q2, Q3, Q4).

**Figure 2**. 1-year predicted lung cancer probabilities from the smoking and integrated risk prediction models in the validation study (EPIC and NSHDS, ever smokers).

The validation samples consist of EPIC and NSHDS ever smoking subjects diagnosed within 1 year of blood collection. For the controls, the size of the points is proportional to the number of eligible subjects represented (corresponding to 1/sampling probability). The right panel represents a zoomed version of the full figure.**Figure 3**. Receiver Operating Characteristics (ROC) curve analysis in the validation study (EPIC and NSHDS, ever smokers).

A: ROC curve analysis in the validation study (EPIC and NSHDS ever smoker subjects diagnosed within 1 year of blood collection) for two risk prediction models, smoking variables only, and an integrated model with the smoking variables and the biomarker score combined. B: Sensitivity and specificity in relation to the 1-year predicted lung cancer probability from the integrated model.